Reading comprehension (RC)

Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + human editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary + cloze	RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Squad: 100,000+ questions for machine comprehension of text (Rajpurkar acl'2016)

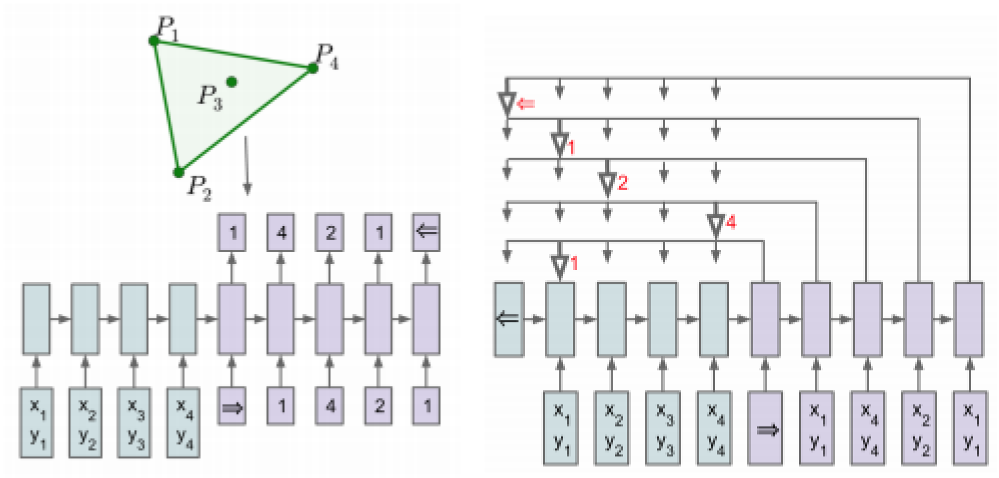
Reading comprehension (RC)

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowd-workers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Squad: 100,000+ questions for machine comprehension of text (Rajpurkar acl'2016)

Pointer networks （Vinyals nips’2015)

Motivation: It deals with the fundamental problem of representing variable length dictionaries by using a softmax probability distribution as a “pointer”.



(a) Sequence-to-Sequence

(b) Ptr-Net

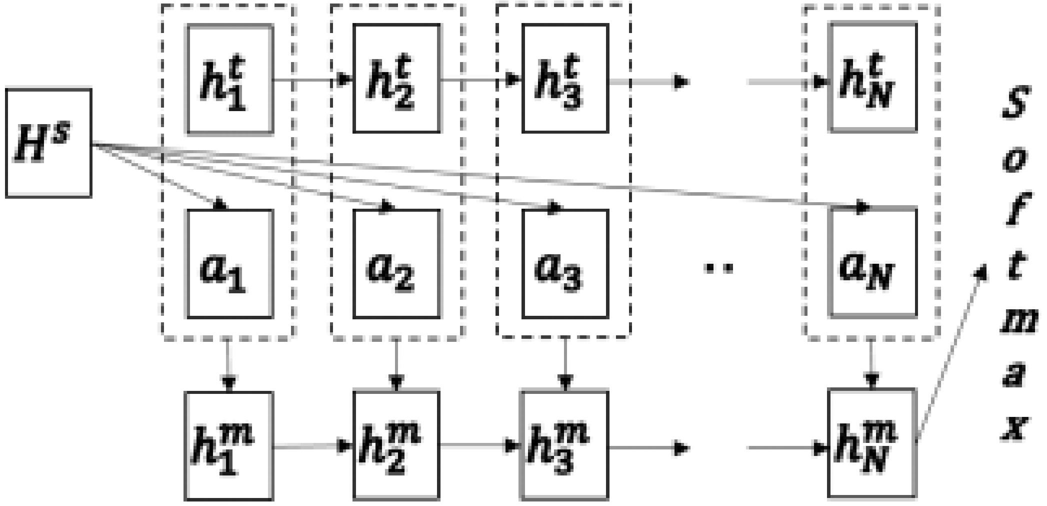
$$
\begin{aligned}
u_j^i &= v^T \tanh(W_1 e_j + W_2 d_i) \quad j \in (1, \ldots, n) \\
p(C_i | C_1, \ldots, C_{i-1}, \mathcal{P}) &= \mathrm{softmax}(u^i)
\end{aligned}
$$

Learning natural language inference with LSTM(Wang NAACL'2016)

Task : Natural language inference (NLI)
Natural language inference (NLI) is the problem of determining whether from a premise sentence P one can infer another hypothesis sentence H (MacCartney, 2009)

https://github.com/shuohangwang/ SeqMatchSeq

H^s && H^t -> entailment, contradiction and neutral



$$\mathbf{a}_k = \sum_{j=1}^{M} \alpha_{kj} \mathbf{h}_j^{\mathrm{s}}, \qquad \alpha_{kj} = \frac{\exp(e_{kj})}{\sum_{j'} \exp(e_{kj'})},$$

$$e_{kj} = \mathbf{w}^{\mathrm{e}} \cdot \tanh(\mathbf{W}^{\mathrm{s}}\mathbf{h}_j^{\mathrm{s}} + \mathbf{W}^{\mathrm{t}}\mathbf{h}_k^{\mathrm{t}} + \mathbf{W}^{\mathrm{m}}\mathbf{h}_{k-1}^{\mathrm{m}}).$$

$$\mathbf{m}_k = \begin{bmatrix} \mathbf{a}_k \\ \mathbf{h}_k^{\mathrm{t}} \end{bmatrix}.$$

$$\begin{aligned}
\mathbf{i}_k^{\mathrm{m}} &= \sigma(\mathbf{W}^{\mathrm{mi}}\mathbf{m}_k + \mathbf{V}^{\mathrm{mi}}\mathbf{h}_{k-1}^{\mathrm{m}} + \mathbf{b}^{\mathrm{mi}}), \\
\mathbf{f}_k^{\mathrm{m}} &= \sigma(\mathbf{W}^{\mathrm{mf}}\mathbf{m}_k + \mathbf{V}^{\mathrm{mf}}\mathbf{h}_{k-1}^{\mathrm{m}} + \mathbf{b}^{\mathrm{mf}}), \\
\mathbf{o}_k^{\mathrm{m}} &= \sigma(\mathbf{W}^{\mathrm{mo}}\mathbf{m}_k + \mathbf{V}^{\mathrm{mo}}\mathbf{h}_{k-1}^{\mathrm{m}} + \mathbf{b}^{\mathrm{mo}}), \\
\mathbf{c}_k^{\mathrm{m}} &= \mathbf{f}_k^{\mathrm{m}} \odot \mathbf{c}_{k-1}^{\mathrm{m}} + \mathbf{i}_k^{\mathrm{m}} \odot \tanh(\mathbf{W}^{\mathrm{mc}}\mathbf{m}_k + \mathbf{V}^{\mathrm{mc}}\mathbf{h}_{k-1}^{\mathrm{m}} \\
&\quad + \mathbf{b}^{\mathrm{mc}}), \\
\mathbf{h}_k^{\mathrm{m}} &= \mathbf{o}_k^{\mathrm{m}} \odot \tanh(\mathbf{c}_k^{\mathrm{m}}). \qquad (8)
\end{aligned}$$

MACHINE COMPREHENSION USING MATCH-LSTM AND ANSWER POINTER(wang ICLR'2017)

Task: reading comprehension(RC)

LSTM Preprocessing Layer

$$\mathbf{H}^{\mathrm{p}} = \overrightarrow{LSTM}(\mathbf{P}), \quad \mathbf{H}^{\mathrm{q}} = \overrightarrow{LSTM}(\mathbf{Q}).$$

Match-LSTM Layer

$$\overrightarrow{\mathbf{G}}_i = \tanh(\mathbf{W}^{\mathrm{q}}\mathbf{H}^{\mathrm{q}} + (\mathbf{W}^{\mathrm{p}}\mathbf{h}_i^{\mathrm{p}} + \mathbf{W}^{\mathrm{r}}\overrightarrow{\mathbf{h}}_{i-1}^{\mathrm{r}} + \mathbf{b}^{\mathrm{p}}) \otimes \mathbf{e}_Q),$$

$$\overrightarrow{\alpha}_i = \mathrm{softmax}(\mathbf{w}^{\intercal}\overrightarrow{\mathbf{G}}_i + b \otimes \mathbf{e}_Q),$$

$$\overrightarrow{\mathbf{z}}_i = \begin{bmatrix} \mathbf{h}_i^{\mathrm{p}} \\ \mathbf{H}^{\mathrm{q}}\overrightarrow{\alpha}_i^{\intercal} \end{bmatrix}, \qquad \overrightarrow{\mathbf{h}}_i^{\mathrm{r}} = \overrightarrow{LSTM}(\overrightarrow{\mathbf{z}}_i, \overrightarrow{\mathbf{h}}_{i-1}^{\mathrm{r}}), \qquad \mathbf{H}^{\mathrm{r}} = \begin{bmatrix} \overrightarrow{\mathbf{H}}^{\mathrm{r}} \\ \overleftarrow{\mathbf{H}}^{\mathrm{r}} \end{bmatrix}$$

Answer Pointer Layer

$$\mathbf{F}_k = \tanh(\mathbf{V}\widetilde{\mathbf{H}}^{\mathrm{r}} + (\mathbf{W}^{\mathrm{a}}\mathbf{h}_{k-1}^{\mathrm{a}} + \mathbf{b}^{a}) \otimes \mathbf{e}_{(P+1)}),$$

$$\beta_k = \mathrm{softmax}(\mathbf{v}^{\intercal}\mathbf{F}_k + c \otimes \mathbf{e}_{(P+1)}),$$

$$\mathbf{h}_k^{\mathrm{a}} = \overrightarrow{LSTM}(\widetilde{\mathbf{H}}^{\mathrm{r}}\beta_k^{\intercal}, \mathbf{h}_{k-1}^{\mathrm{a}}).$$

$$p(\mathbf{a}|\mathbf{H}^{\mathrm{r}}) = \prod_k p(a_k|a_1, a_2, \ldots, a_{k-1}, \mathbf{H}^{\mathrm{r}}),$$

$$p(a_k = j|a_1, a_2, \ldots, a_{k-1}, \mathbf{H}^{\mathrm{r}}) = \beta_{k,j}.$$

MACHINE COMPREHENSION USING MATCH-LSTM AND ANSWER POINTER(wang ICLR’2017)

Task: reading comprehension(RC)

	l	$\|\theta\|$	Exact Match		F1	
			Dev	Test	Dev	Test
Random Guess	-	0	1.1	1.3	4.1	4.3
Logistic Regression	-	-	40.0	40.4	51.0	51.0
DCR	-	-	62.5	62.5	71.2	71.0
Match-LSTM with Ans-Ptr (Sequence)	150	882K	54.4	-	68.2	-
Match-LSTM with Ans-Ptr (Boundary)	150	882K	61.1	-	71.2	-
Match-LSTM with Ans-Ptr (Boundary+Search)	150	882K	63.0	-	72.7	-
Match-LSTM with Ans-Ptr (Boundary+Search)	300	3.2M	63.1	-	72.7	-
Match-LSTM with Ans-Ptr (Boundary+Search+b)	150	1.1M	63.4	-	73.0	-
Match-LSTM with Bi-Ans-Ptr (Boundary+Search+b)	150	1.4M	**64.1**	**64.7**	**73.9**	**73.7**
Match-LSTM with Ans-Ptr (Boundary+Search+en)	150	882K	**67.6**	**67.9**	**76.8**	**77.0**

Table 2: Experiment Results. Here “Search” refers to globally searching the spans with no more than 15 tokens, “b” refers to using bi-directional pre-processing LSTM, and “en” refers to ensemble method.

R-net (msra,2017)

Task: reading comprehension(RC)

Question and Passage Encoder	$u_t^Q = \mathrm{BiRNN}_Q(u_{t-1}^Q, [e_t^Q, c_t^Q]) \quad (1)$ $u_t^P = \mathrm{BiRNN}_P(u_{t-1}^P, [e_t^P, c_t^P]) \quad (2)$
Gated Attention-based Recurrent Networks	$v_t^P = \mathrm{RNN}(v_{t-1}^P, [u_t^P, c_t])$ $g_t = \mathrm{sigmoid}(W_g[u_t^P, c_t])$ $[u_t^P, c_t]^* = g_t \odot [u_t^P, c_t] \quad (6)$ $s_j^t = \mathrm{v}^{\mathrm{T}}\tanh(W_u^Q u_j^Q + W_u^P u_t^P + W_v^P v_{t-1}^P)$ $a_i^t = \exp(s_i^t)/\Sigma_{j=1}^m \exp(s_j^t)$ $c_t = \Sigma_{i=1}^m a_i^t u_i^Q \quad (4)$

R-net (msra,2017)

Task: reading comprehension(RC)

Self-Matching Attention	$$h_t^P = \mathrm{BiRNN}(h_{t-1}^P, [v_t^P, c_t]) \quad (7)$$ $$c_t = att(v^P, v_t^P)$$	$$s_j^t = \mathrm{v}^{\mathrm{T}}\tanh(W_v^P v_j^P + W_v^{\tilde{P}} v_t^P)$$ $$a_i^t = \exp(s_i^t)/\Sigma_{j=1}^n \exp(s_j^t)$$ $$c_t = \Sigma_{i=1}^n a_i^t v_i^P \quad (8)$$
Output Layer	$$s_j^t = \mathrm{v}^{\mathrm{T}}\tanh(W_h^P h_j^P + W_{\mathrm{h}}^a h_{t-1}^a)$$ $$a_i^t = \exp(s_i^t)/\Sigma_{j=1}^n \exp(s_j^t)$$ $$p^t = \arg\max(a_1^t, \ldots, a_n^t) \quad (9)$$	$$c_t = \Sigma_{i=1}^n a_i^t h_i^P$$ $$h_t^a = \mathrm{RNN}(h_{t-1}^a, c_t) \quad (10)$$

Output layers initial hidden state

$$r^Q = att(u^Q, V_r^Q)$$

$$s_j = \mathrm{v}^{\mathrm{T}}\tanh(W_u^Q u_j^Q + W_{\mathrm{v}}^Q V_r^Q)$$

$$a_i = \exp(s_i)/\Sigma_{j=1}^m \exp(s_j)$$

$$r^Q = \Sigma_{i=1}^m a_i u_i^Q$$

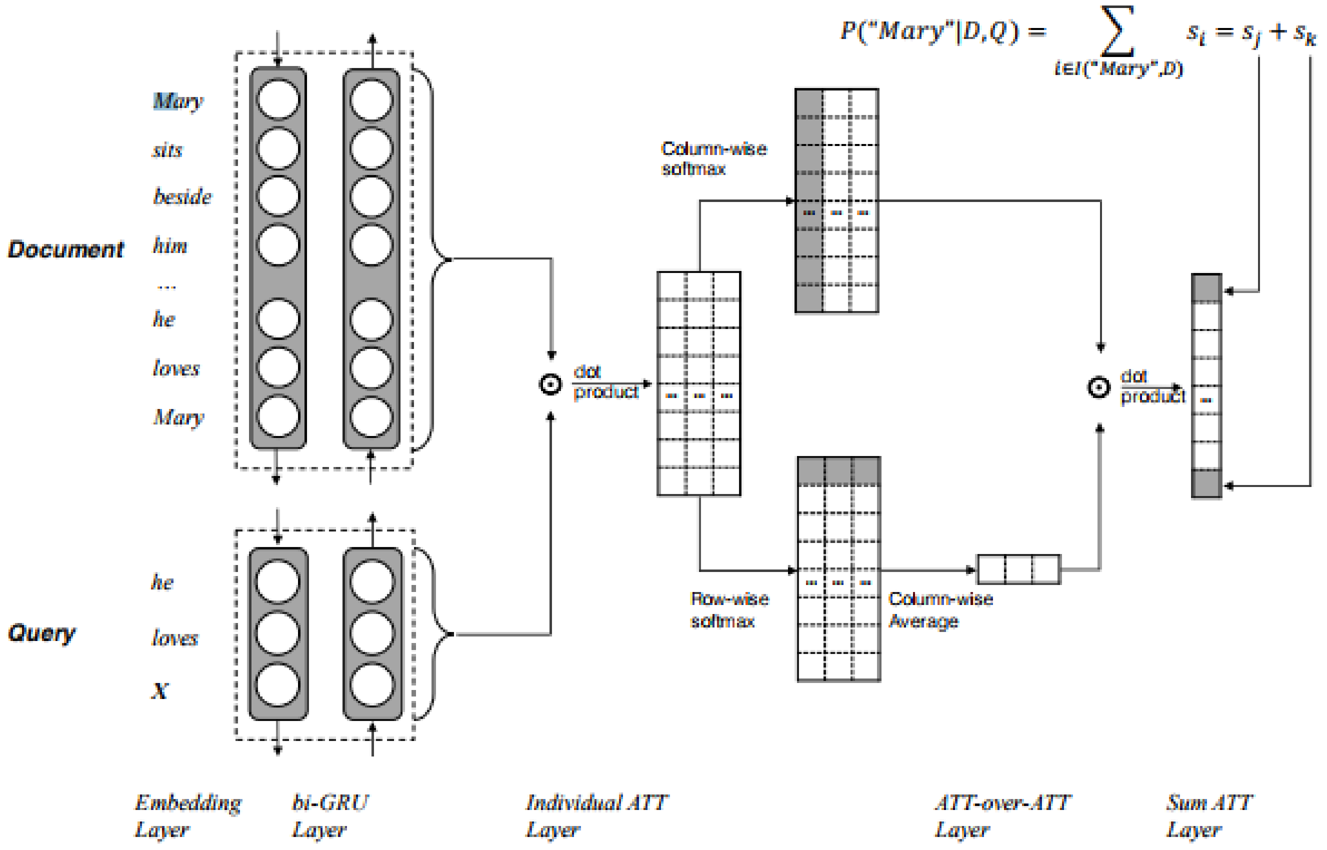
R-net (msra,2017)

Task: reading comprehension(RC)

	Dev Set	Test Set
Single model	**EM / F1**	**EM / F1**
LR Baseline (Rajpurkar et al., 2016)	40.0 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.2	62.5 / 71.0
Attentive CNN context with LSTM (NLPR, CASIA)	- / -	63.3 / 73.5
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016b)	64.1 / 73.9	64.7 / 73.7
Dynamic Coattention Networks (Xiong et al., 2016)	65.4 / 75.6	66.2 / 75.9
Iterative Coattention Network (Fudan University)	- / -	67.5 / 76.8
FastQA (Weissenborn et al., 2017)	- / -	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
T-gating (Peking University)	- / -	68.1 / 77.6
RaSoR (Lee et al., 2016)	- / -	69.6 / 77.7
SEDT+BiDAF (Liu et al., 2017)	- / -	68.5 / 78.0
Multi-Perspective Matching (Wang et al., 2016)	- / -	70.4 / 78.8
FastQAExt (Weissenborn et al., 2017)	- / -	70.8 / 78.9
Mnemonic Reader (NUDT & Fudan University)	- / -	69.9 / 79.2
Document Reader (Chen et al., 2017)	- / -	70.7 / 79.4
ReasoNet (Shen et al., 2016)	- / -	70.6 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	- / -	70.6 / 79.5
jNet (Zhang et al., 2017)	- / -	70.6 / 79.8
Interactive AoA Reader (Joint Laboratory of HIT and iFLYTEK Research)	- / -	71.2 / 79.9
R-NET (Wang et al., 2017)	**71.1 / 79.5**	**71.3 / 79.7**
R-NET (March 2017)	**72.3 / 80.6**	**72.3 / 80.7**

AoA reader (hit,2016)

Task: factoid reading comprehension(RC)



Multiperspective matching (wang,2016)

Task: reading comprehension(RC)

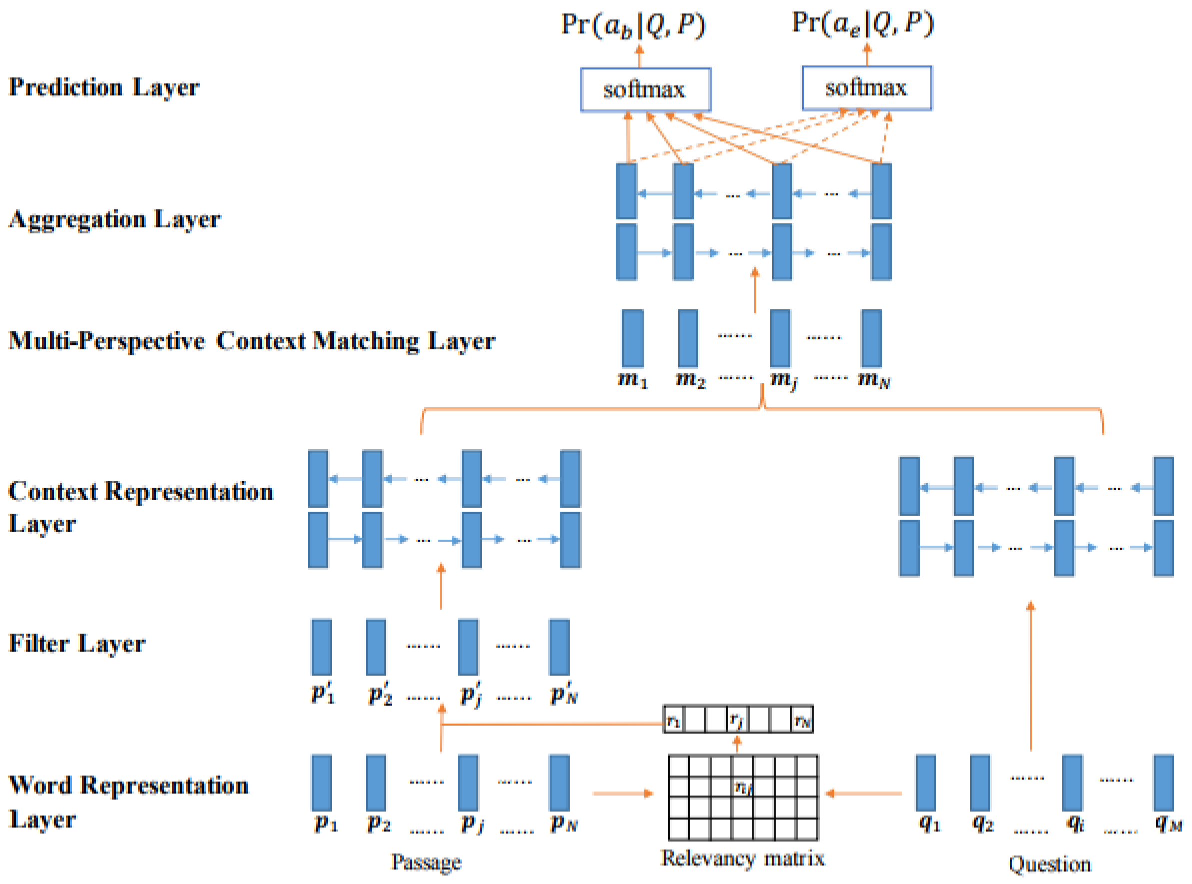


Figure 1: Architecture for Multi-Perspective Context Matching Model.

$$\boldsymbol{m} = f_m(\boldsymbol{v}_1, \boldsymbol{v}_2; \boldsymbol{W}) \tag{5}$$

$$\boldsymbol{m} = [m_1, ..., m_k, ..., m_l]$$

$$m_k = cosine(W_k \circ \boldsymbol{v}_1, W_k \circ \boldsymbol{v}_2)$$

$$\begin{aligned} \overrightarrow{\boldsymbol{m}}_j^{full} &= f_m(\overrightarrow{\boldsymbol{h}}_j^p, \overrightarrow{\boldsymbol{h}}_M^q; \boldsymbol{W}^1) \\ \overleftarrow{\boldsymbol{m}}_j^{full} &= f_m(\overleftarrow{\boldsymbol{h}}_j^p, \overleftarrow{\boldsymbol{h}}_1^q; \boldsymbol{W}^2) \end{aligned} \tag{7}$$

$$\begin{aligned} \overrightarrow{\boldsymbol{m}}_j^{max} &= \max_{i \in (1...M)} f_m(\overrightarrow{\boldsymbol{h}}_j^p, \overrightarrow{\boldsymbol{h}}_i^q; \boldsymbol{W}^3) \\ \overleftarrow{\boldsymbol{m}}_j^{max} &= \max_{i \in (1...M)} f_m(\overleftarrow{\boldsymbol{h}}_j^p, \overleftarrow{\boldsymbol{h}}_i^q; \boldsymbol{W}^4) \end{aligned} \tag{8}$$

$$\begin{aligned} \overrightarrow{\boldsymbol{m}}_j^{mean} &= \frac{1}{M} \sum_{i=1}^{M} f_m(\overrightarrow{\boldsymbol{h}}_j^p, \overrightarrow{\boldsymbol{h}}_i^q; \boldsymbol{W}^5) \\ \overleftarrow{\boldsymbol{m}}_j^{mean} &= \frac{1}{M} \sum_{i=1}^{M} f_m(\overleftarrow{\boldsymbol{h}}_j^p, \overleftarrow{\boldsymbol{h}}_i^q; \boldsymbol{W}^6) \end{aligned} \tag{9}$$

Reading Wikipedia to Answer Open-Domain Questions(chen,ACL'2017)

Task: retrieval and reading comprehension(RC)

Document Reader $\{q_1, \ldots, q_l\}$ $\{p_1, \ldots, p_m\}$

Paragraph encoding

Word embeddings: $f_{emb}(p_i) = \mathbf{E}(p_i)$

Exact match: $f_{exact\_match}(p_i) = \mathbb{I}(p_i \in q)$

Token features: $f_{token}(p_i) = (\mathrm{POS}(p_i), \mathrm{NER}(p_i), \mathrm{TF}(p_i))$

Aligned question embedding $\sum_j a_{i,j}\mathbf{E}(q_j)$ $a_{i,j} = \frac{\exp\left(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j))\right)}{\sum_{j'} \exp\left(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_{j'}))\right)}$

$$\{\mathbf{p}_1, \ldots, \mathbf{p}_m\} = \mathrm{RNN}(\{\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_m\})$$

Question encoding $\mathbf{q} = \sum_j b_j \mathbf{q}_j$ $b_j = \frac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})}$

Reading Wikipedia to Answer Open-Domain Questions(chen,ACL'2017)

Task: retrieval and reading comprehension(RC)

Document Reader

Prediction

$$P_{start}(i) \propto \exp(\mathbf{p}_i \mathbf{W}_s \mathbf{q})$$
$$P_{end}(i) \propto \exp(\mathbf{p}_i \mathbf{W}_e \mathbf{q})$$

R-net†	n/a	n/a	71.3	79.7
DrQA (Our model, Document Reader Only)	**69.5**	**78.8**	70.0	79.0

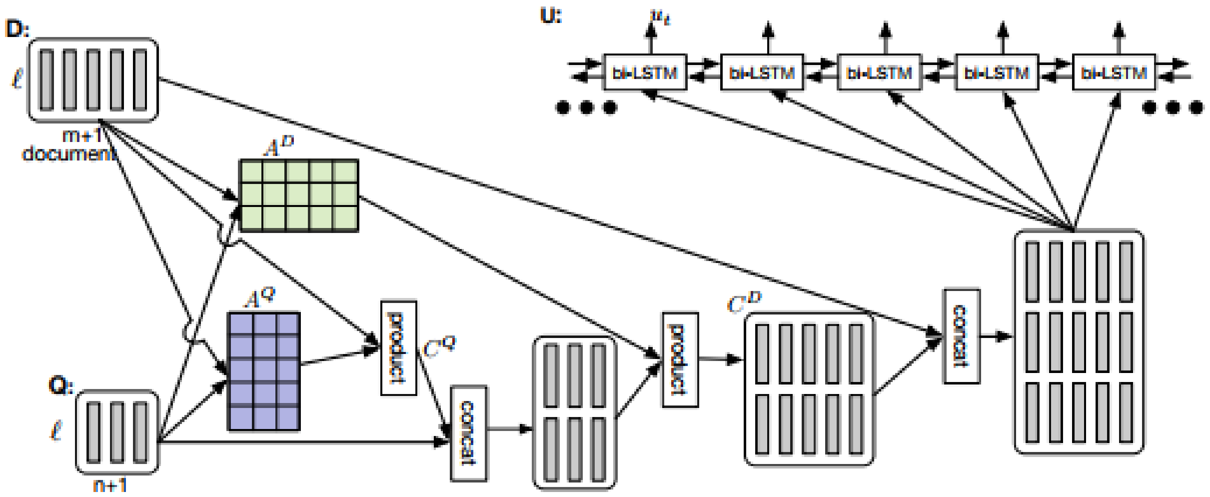
For Cloze test

$$\alpha_i = \mathrm{softmax}_i \, \mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i$$
$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

$$a = \arg\max_{a \in p \cap E} W_a^\intercal \mathbf{o}$$

DYNAMIC COATTENTION NETWORKS FOR QUESTION ANSWERING(saleforce,iclr2017)

Task: reading comprehension(RC)



$$L = D^\top Q \in \mathbb{R}^{(m+1)\times(n+1)}$$

$$A^Q = \mathrm{softmax}(L) \in \mathbb{R}^{(m+1)\times(n+1)} \text{ and } A^D = \mathrm{softmax}\left(L^\top\right) \in \mathbb{R}^{(n+1)\times(m+1)}$$

$$C^Q = DA^Q \in \mathbb{R}^{\ell\times(n+1)}.$$

$$C^D = \left[Q; C^Q\right] A^D \in \mathbb{R}^{2\ell\times(m+1)}.$$

$$u_t = \text{Bi-LSTM}\left(u_{t-1}, u_{t+1}, \left[d_t; c_t^D\right]\right) \in \mathbb{R}^{2\ell}.$$

DYNAMIC COATTENTION NETWORKS FOR QUESTION ANSWERING(saleforce,iclr2017)

Task: reading comprehension(RC)

DYNAMIC POINTING DECODER

$$h_i = \text{LSTM}_{dec}\left(h_{i-1}, \left[u_{s_{i-1}}; u_{e_{i-1}}\right]\right)$$

$$s_i = \underset{t}{\text{argmax}}\left(\alpha_1, \ldots, \alpha_m\right)$$

$$e_i = \underset{t}{\text{argmax}}\left(\beta_1, \ldots, \beta_m\right)$$

$$\alpha_t = \text{HMN}_{start}\left(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}\right)$$

$$\text{HMN}\left(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}\right) = \max\left(W^{(3)}\left[m_t^{(1)}; m_t^{(2)}\right] + b^{(3)}\right) \tag{9}$$

$$r = \tanh\left(W^{(D)}\left[h_i; u_{s_{i-1}}; u_{e_{i-1}}\right]\right) \tag{10}$$

$$m_t^{(1)} = \max\left(W^{(1)}\left[u_t; r\right] + b^{(1)}\right) \tag{11}$$

$$m_t^{(2)} = \max\left(W^{(2)} m_t^{(1)} + b^{(2)}\right) \tag{12}$$

DYNAMIC COATTENTION NETWORKS FOR QUESTION ANSWERING(saleforce,iclr2017)

Task: reading comprehension(RC)

DYNAMIC POINTING DECODER

$$h_i = \text{LSTM}_{dec}\left(h_{i-1}, \left[u_{s_{i-1}}; u_{e_{i-1}}\right]\right)$$

$$s_i = \underset{t}{\text{argmax}}\left(\alpha_1, \ldots, \alpha_m\right)$$

$$e_i = \underset{t}{\text{argmax}}\left(\beta_1, \ldots, \beta_m\right)$$

$$\alpha_t = \text{HMN}_{start}\left(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}\right)$$

$$\text{HMN}\left(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}\right) = \max\left(W^{(3)}\left[m_t^{(1)}; m_t^{(2)}\right] + b^{(3)}\right) \tag{9}$$

$$r = \tanh\left(W^{(D)}\left[h_i; u_{s_{i-1}}; u_{e_{i-1}}\right]\right) \tag{10}$$

$$m_t^{(1)} = \max\left(W^{(1)}\left[u_t; r\right] + b^{(1)}\right) \tag{11}$$

$$m_t^{(2)} = \max\left(W^{(2)} m_t^{(1)} + b^{(2)}\right) \tag{12}$$

Words or Characters? Fine-grained Gating for Reading Comprehension(2017'ICLR)

Task: reading comprehension(RC)