Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification

Yizhong Wang[1 *], Kai Liu[2], Jing Liu[2], Wei He[2],
Yajuan Lyu[2], Hua Wu[2], Sujian Li[1] and Haifeng Wang[2]

[1]Key Laboratory of Computational Linguistics, Peking University, MOE, China
[2]Baidu Inc., Beijing, China
{yizhong, lisujian}@pku.edu.cn, {liukai20, liujing46,
hewei06, lvyajuan, wu_hua, wanghaifeng}@baidu.com

Wang Y, Liu K, Liu J, et al. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018, 1: 1918-1927.

Datasets

1. Cloze style
2. Multiple choice
3. Text span
   - MS MARCO
   - DuReader

- **MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.** Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. arXiv preprint arXiv:1611.09268 (2016).
- **DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications.** Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. ACL 2018 Workshop.

Datasets

Dataset	Lang	#Que.	#Docs	Source of Que.	Source of Docs	Answer Type
CNN/DM (Hermann et al., 2015)	EN	1.4M	300K	Synthetic cloze	News	Fill in entity
HLF-RC (Cui et al., 2016)	ZH	100K	28K	Synthetic cloze	Fairy/News	Fill in word
CBT (Hill et al., 2015)	EN	688K	108	Synthetic cloze	Children's books	Multi. choices
RACE (Lai et al., 2017)	EN	870K	50K	English exam	English exam	Multi. choices
MCTest (Richardson et al., 2013)	EN	2K	500	Crowdsourced	Fictional stories	Multi. choices
NewsQA (Trischler et al., 2017)	EN	100K	10K	Crowdsourced	CNN	Span of words
SQuAD (Rajpurkar et al., 2016)	EN	100K	536	Crowdsourced	Wiki.	Span of words
SearchQA (Dunn et al., 2017)	EN	140K	6.9M	QA site	Web doc.	Span of words
TrivaQA (Joshi et al., 2017)	EN	40K	660K	Trivia websites	Wiki./Web doc.	Span/substring of words
NarrativeQA (Kočiskỳ et al., 2017)	EN	46K	1.5K	Crowdsourced	Book&movie	Manual summary
MS-MARCO (Nguyen et al., 2016)	EN	100K	200K[1]	User logs	Web doc.	Manual summary
DuReader (this paper)	**ZH**	**200k**	**1M**	**User logs**	**Web doc./CQA**	**Manual summary**

- **MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.** Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. arXiv preprint arXiv:1611.09268 (2016).
- **DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications.** Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. ACL 2018 Workshop.

MS MARCO

Question: What is the difference between a mixed and pure culture?

Passages:

[1] **A culture is a society's total way of living and a society is a group that live in a defined territory and participate in common culture.** While the answer given is in essence true, societies originally form for the express purpose to enhance . . .

[2] . . . There has been resurgence in the economic system known as capitalism during the past two decades. 4. **The mixed economy is a balance between socialism and capitalism.** As a result, some institutions are owned and maintained by . . .

[3] **A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.** Culture on the other hand, is the lifestyle that the people in the country . . .

[4] Best Answer: **A pure culture comprises a single species or strains. A mixed culture is taken from a source and may contain multiple strains or species.** A contaminated culture contains organisms that derived from some place . . .

[5] . . . It will be at that time when we can truly obtain a pure culture. **A pure culture is a culture consisting of only one strain.** You can obtain a pure culture by picking out a small portion of the mixed culture . . .

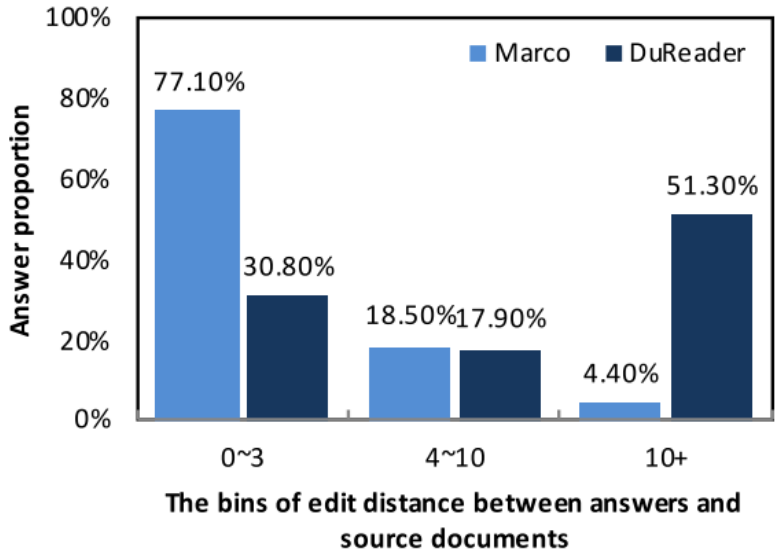
[6] **A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.** A pure culture is a culture consisting of only one strain. . . .

. . . . . .

Reference Answer: A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.

DuReader

- Baidu Search
- Baidu Zhidao



问题：	小米 6 防水等级
答案：	1. 小米 6 是支持 IP68 级的防水。 2. IP56 吧。
文本 1：	**标题：** 小米 6 防水吗小米 6 的防水级别是多少?有问必答_安卓中文网
	段落： 1. …防水作为目前高端手机的标配,特别是苹果也支持防水之后,国产大多数高端旗舰手机都已经支持防水…根据评测资料显示,小米 6 是支持 IP68 级的防水… 2. 下面我们一起来看看,小米 6 的 IP68 级别防水,能达到什么样的效果和功能。我们查询资料得知,IPXX,其中 XX 为两个阿拉伯数字,第一标记数字表示接触保护和外来物保护等级,第二标记数字表示防水保护等级,数字越大表示其防护等级越好。 第一个 X 表示防尘等级,不同数字代表不同的防尘级别,具体如下…如果说小米的防水等级是 IP68,那么表示小米 6 能够在一定的水深环境下,长时间浸水… 3. 提示:支持键盘“←→”键翻页 阅读全文
	是否被选中： 是
…	…
文本 5：	**标题：** 小米 6 防水等级多少能浸泡吗 - 小米社区官方论坛
	段落： 1. 扫描二维码,手机查看本帖 2. 总评分: 经验 +1
	是否被选中： 否
问题类型：	ENTITY
事实/观点：	事实
问题 ID：	181597

Question and Passage Modeling

1. Map each word into the vector space by concatenating its word embedding and sum of its character embeddings.

2. Employ bi-directional LSTMs (BiLSTM) to encode the question Q and passages P_i.

$$\mathbf{u}_t^Q = \mathrm{BiLSTM}_Q(\mathbf{u}_{t-1}^Q, [\mathbf{e}_t^Q, \mathbf{c}_t^Q]) \qquad (1)$$

$$\mathbf{u}_t^{P_i} = \mathrm{BiLSTM}_P(\mathbf{u}_{t-1}^{P_i}, [\mathbf{e}_t^{P_i}, \mathbf{c}_t^{P_i}]) \qquad (2)$$

Question and Passage Modeling

3. The similarity between the t^{th}word in the question and the k^{th}word in passage i is computed as:

$$\mathbf{S}_{t,k} = \mathbf{u}_t^{Q\intercal} \cdot \mathbf{u}_k^{P_i} \qquad (3)$$

4. Follow [1] to get question-aware passage representation $\tilde{\mathbf{u}}_t^{P_i}$. Fuse the contextual information and get the new representation for each word in the passage.

$$\mathbf{v}_t^{P_i} = \mathrm{BiLSTM}_M(\mathbf{v}_{t-1}^{P_i}, \tilde{\mathbf{u}}_t^{P_i}) \qquad (4)$$

---

1. Seo M Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension[J]. 2016 arXiv preprint arXiv:1611.01603.

Answer Boundary Prediction

k^{th} word in the passage to be

The start position probability: α_k^1
The end position probability: α_k^2

$$g_k^t = \mathbf{w}_1^{a\intercal} \tanh(\mathbf{W}_2^a[\mathbf{v}_k^P, \mathbf{h}_{t-1}^a]) \quad (5)$$

$$\alpha_k^t = \exp(g_k^t) / \sum_{j=1}^{|\mathbf{P}|} \exp(g_j^t) \quad (6)$$

$$\mathbf{c}_t = \sum_{k=1}^{|\mathbf{P}|} \alpha_k^t \mathbf{v}_k^P \quad (7)$$

$$\mathbf{h}_t^a = \mathrm{LSTM}(\mathbf{h}_{t-1}^a, \mathbf{c}_t) \quad (8)$$

Train by minimizing the negative log probabilities of the true start and end indices.

$$\mathcal{L}_{boundary} = -\frac{1}{N} \sum_{i=1}^{N} (\log \alpha_{y_i^1}^1 + \log \alpha_{y_i^2}^2) \quad (9)$$

---

1. Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer.arXiv preprint arXiv:1608.07905 .

Answer Content Modeling

1. Predict whether each word should be included in the content of the answer. The k^{th} word's content probability:

$$p_k^c = \text{sigmoid}(\mathbf{w}_1^{c\intercal}\text{ReLU}(\mathbf{W}_2^c\mathbf{v}_k^{P_i})) \qquad (10)$$

The words within the answer span is labeled as 1 and other words is labeled as 0.

$$\mathcal{L}_{content} = -\frac{1}{N}\frac{1}{|\mathrm{P}|}\sum_{i=1}^{N}\sum_{j=1}^{|P|}[y_k^c \log p_k^c + (1-y_k^c)\log(1-p_k^c)] \qquad (11)$$

2. The answer from passage i is represented as a weighted sum of all the word embeddings in this passage:

$$\mathbf{r}^{A_i} = \frac{1}{|\mathbf{P}_i|}\sum_{k=1}^{|\mathbf{P}_i|} p_k^c[\mathbf{e}_k^{P_i}, \mathbf{c}_k^{P_i}] \qquad (12)$$

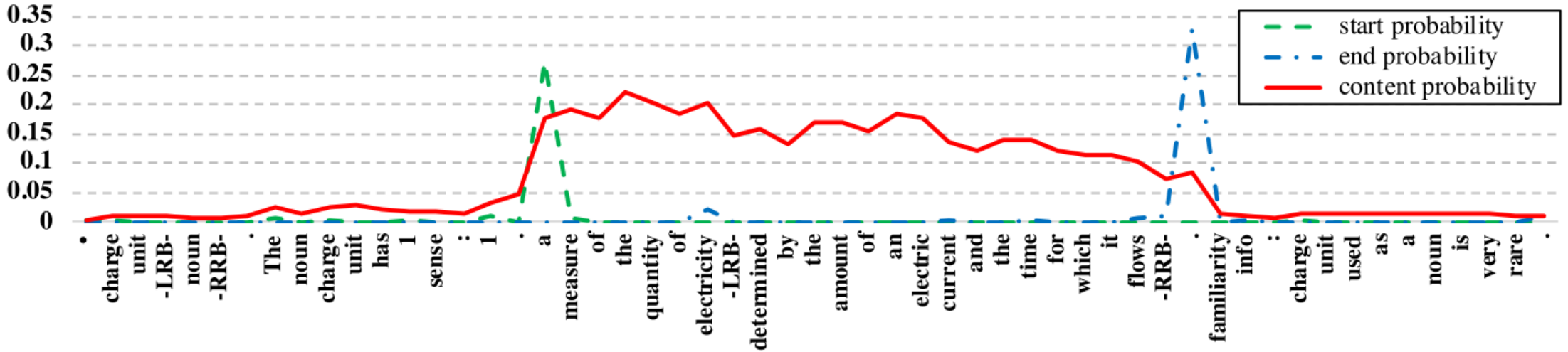


Figure 2: The boundary probabilities and content probabilities for the words in a passage

Charge unit LRB noun RRB. The noun charge unit has 1 sense: 1. **a measure of the quantity of electricity LRB determined by the amount of an electric current and the time for which it flows RRB.** Familiarity info : charge unit used as a noun is very rare.

Cross-Passage Answer Verification

1. Collect supportive information via attention mechanism:

$$s_{i,j} = \begin{cases} 0, & \text{if } i = j, \\ \mathbf{r}^{A_i\intercal} \cdot \mathbf{r}^{A_j}, & \text{otherwise} \end{cases} \tag{13}$$

$$\alpha_{i,j} = \exp(s_{i,j}) / \sum_{k=1}^{n} \exp(s_{i,k}) \tag{14}$$

$$\tilde{\mathbf{r}}^{A_i} = \sum_{j=1}^{n} \alpha_{i,j} \mathbf{r}^{A_j} \tag{15}$$

2. Pass it together with the original representation $\mathbf{r}^{A_i}$ to a fully connected layer:

$$g_i^v = \mathbf{w}^{v\intercal}[\mathbf{r}^{A_i}, \tilde{\mathbf{r}}^{A_i}, \mathbf{r}^{A_i} \odot \tilde{\mathbf{r}}^{A_i}] \tag{16}$$

Cross-Passage Answer Verification

3. normalize these scores over all passages to get the verification score for answer candidate A_i:

$$p_i^v = \exp(g_i^v) / \sum_{j=1}^{n} \exp(g_j^v) \qquad (17)$$

4. Take the answer from the gold passage as the gold answer.

$$\mathcal{L}_{verify} = -\frac{1}{N} \sum_{i=1}^{N} \log p_{y_i^v}^v \qquad (18)$$

y_i^v is the index of the correct answer in all the answer candidates of the i^{th}instance .

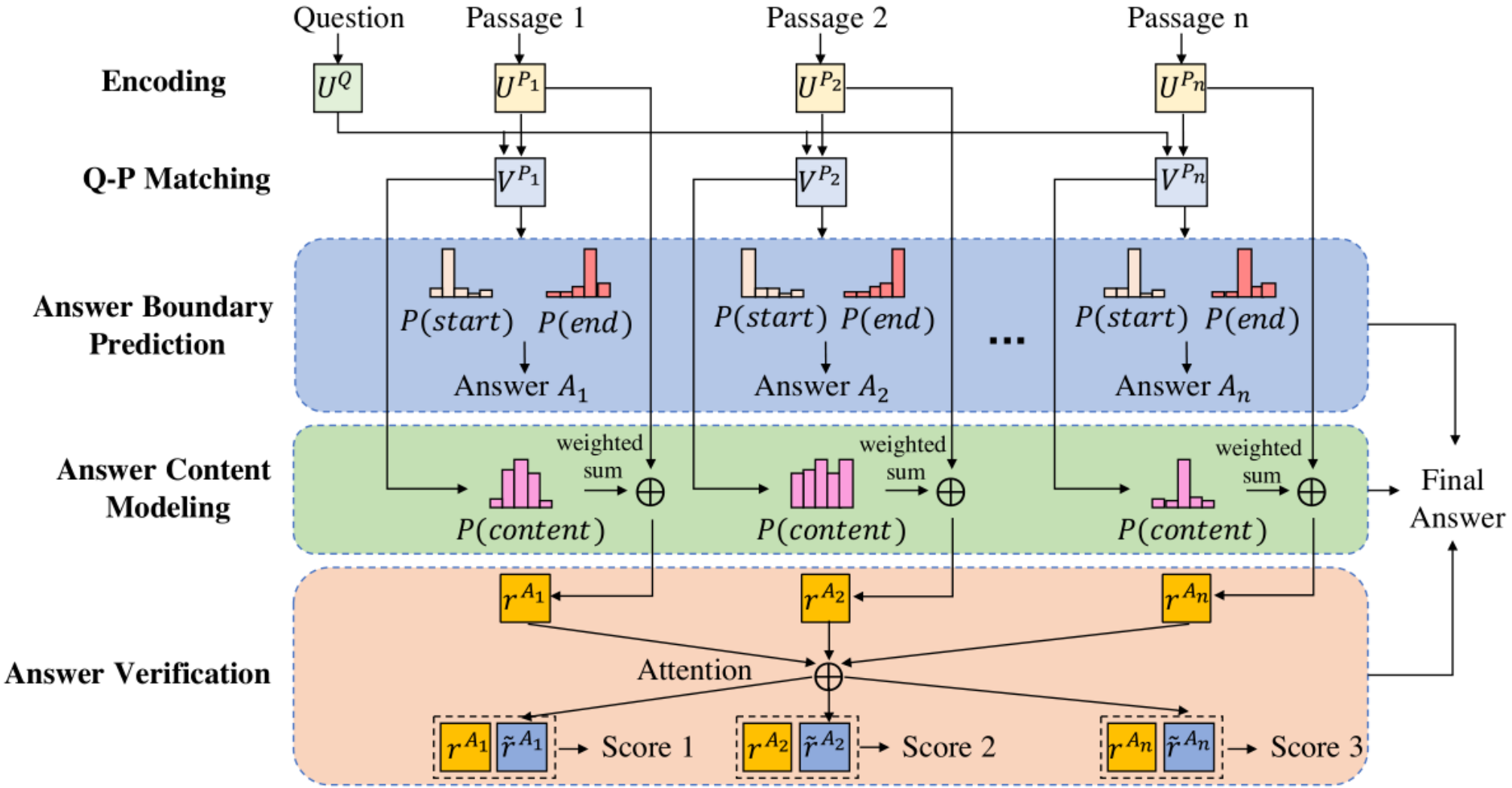


Figure 1: Overview of our method for multi-passage machine reading comprehension

Training and Predicting

1. finding the boundary of the answer;
2. predicting whether each word should be included in the content;
3. selecting the best answer via cross-passage answer verification.

$$\mathcal{L} = \mathcal{L}_{boundary} + \beta_1 \mathcal{L}_{content} + \beta_2 \mathcal{L}_{verify} \quad (19)$$

---

1. Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 .

Training and Predicting

Question: What is the difference between a mixed and pure culture	**Scores**		
Answer Candidates:	**Boundary**	**Content**	**Verification**
[1] A culture is a society's total way of living and a society is a group ...	$\mathbf{1.0 \times 10^{-2}}$	$\mathbf{1.0 \times 10^{-1}}$	1.1×10^{-1}
[2] The mixed economy is a balance between socialism and capitalism.	1.0×10^{-4}	4.0×10^{-2}	3.2×10^{-2}
[3] A pure culture is one in which only one kind of microbial species is ...	5.5×10^{-3}	7.7×10^{-2}	1.2×10^{-1}
[4] A pure culture comprises a single species or strains. A mixed ...	2.7×10^{-3}	8.1×10^{-2}	1.3×10^{-1}
[5] A pure culture is a culture consisting of only one strain.	5.8×10^{-4}	7.9×10^{-2}	5.1×10^{-2}
[6] A pure culture is one in which only one kind of microbial species ...	5.8×10^{-3}	9.1×10^{-2}	$\mathbf{2.7 \times 10^{-1}}$
.....		.....	

Table 6: Scores predicted by our model for the answer candidates shown in Table 1. Although the candidate [1] gets high boundary and content scores, the correct answer [6] is preferred by the verification model and is chosen as the final answer.

Results

Model	ROUGE-L	BLEU-1
FastQA_Ext (Weissenborn et al., 2017)	33.67	33.93
Prediction (Wang and Jiang, 2016)	37.33	40.72
ReasoNet (Shen et al., 2017)	38.81	39.86
R-Net (Wang et al., 2017c)	42.89	42.22
S-Net (Tan et al., 2017)	45.23	43.78
Our Model	**46.15**	**44.47**
S-Net (Ensemble)	46.65	44.78
Our Model (Ensemble)	**46.66**	**45.41**
Human	47	46

Table 3: Performance of our method and competing models on the MS-MARCO test set

Results

Model	BLEU-4	ROUGE-L
Match-LSTM	31.8	39.0
BiDAF	31.9	39.2
PR + BiDAF	37.55	41.81
Our Model	**40.97**	**44.18**
Human	56.1	57.4

Table 4: Performance on the DuReader test set