Carnegie Mellon University
Language Technologies Institute

Nerual QA model on MS-MARCO and DuReader

Zihua Liu, Huiming Jin, Hui Luo, Yihui Peng

Motivation

- Real-world generated questions
- Questions in different domains and lingustic diversity
- Reliable answer: All questions have an answer written by human
- Large-scale data for training complex neural models
- Build a general QA system on multiple datasets in different languages

MS-MARCO Version 2.1

- Version 2.1: 1,010,916 unique real queries generated from Bing usage logs.
- 10 most relevant passages for each question.

Question contains	
YesNo	7.46%
What	34.96%
How	16.8%
Where	3.46%
When	2.71%
Why	1.67%
Who	3.33%
Which	1.79%
Other	27.83%

Question classification	
Description	53.12%
Numeric	26.12%
Entity	8.81%
Location	6.17%
Person	5.78%

From Payal et al. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset

MS-MARCO Version 2.1


```
"answers":["A corporation is a company or group of people authorized to act as
a single entity and recognized as such in law."],
"passages":[
        {
                "is_selected":0,
                "url":"http:\/\/www.wisegeek.com\/what-is-a-corporation.htm",
                "passage_text":"A company is ... or indirectly."},
                ...
        }],
"query":". what is a corporation?",
"query_id":1102432,
"query_type":"DESCRIPTION",
"wellFormedAnswers":"[]"
```

DuReader Version 2.0

- Largest Chinese MRC dataset
- 300K questions, 1.4M documents, 660K human-summarized answers
- Document collected from Baidu Search and Baidu Zhidao.

	Fact	Opinion	Total
Entity	14.4%	13.8%	28.2%
Description	42.8%	21.0%	63.8%
YesNo	2.9%	5.1%	8.0%
Total	60.1%	39.9%	100.0%

From Wei et al. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications

DuReader

	Fact	Opinion
Entity	iphone哪天发布 On which day will iphone be released	2017最好看的十部电影 Top 10 movies of 2017
Description	消防车为什么是红的 Why are firetrucks red	丰田卡罗拉怎么样 How is Toyota Carola
YesNo	39.5度算高烧吗 Is 39.5 degree a high fever	学围棋能开发智力吗 Does learning to play go improve intelligence

From Wei et al. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications

Baseline Models

- BiDAF

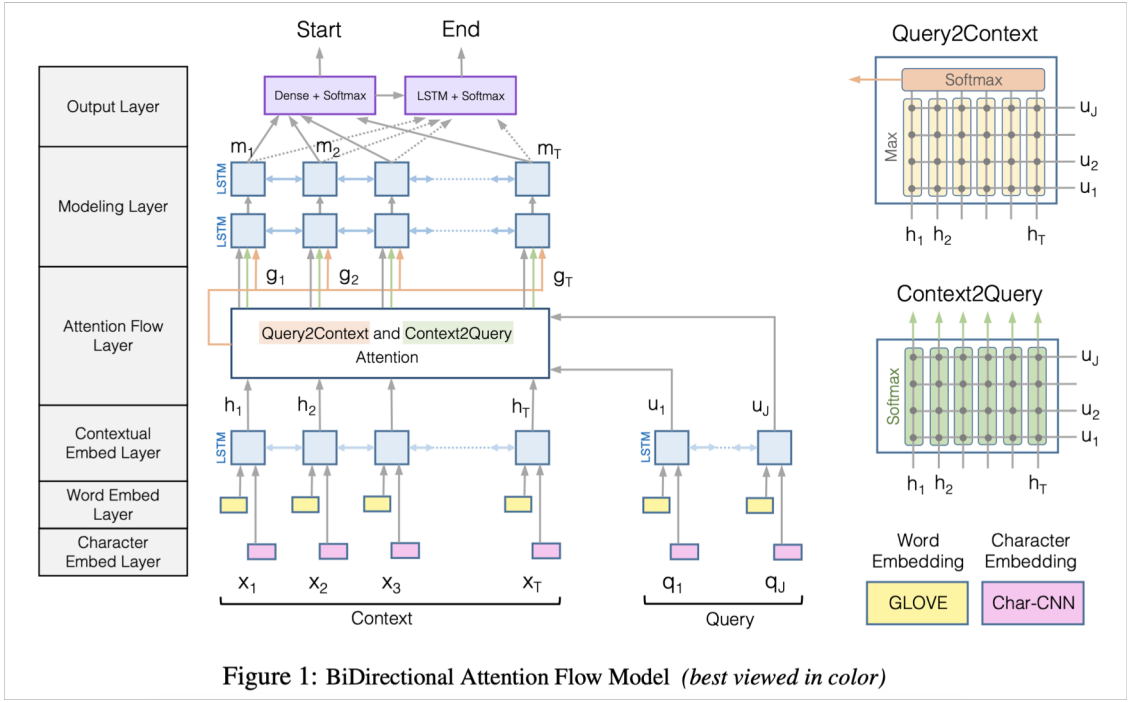
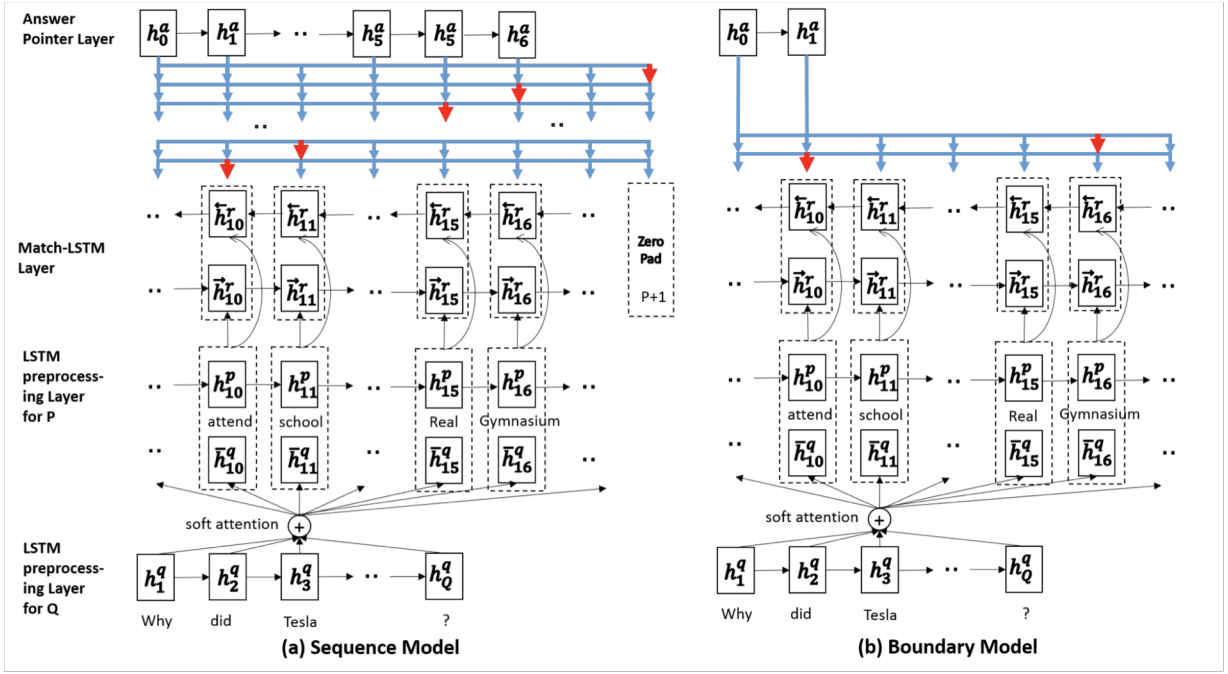


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Baseline Models

- Match-LSTM



Proposed Ideas

- Build a general model using Transfer Learning regardless of the language
- Apply pretrained BERT for word embedding to improve the performance
- Apply cross passage answer verification
- Add an answer boundary prediction module to decrease the search space
- ...

Evaluation Metrics - ROUGE & BLEU Score

MS MARCO | Home | Dataset | **Leaderboard** | Submission | About | Contact us

Q&A Task

Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**Human Performance**	April 23th, 2018	53.870	48.50
2	**Masque Q&A Style** NTT Media Intelligence Laboratories [Nishida et al. '19]	January 3rd, 2019	52.20	43.77
3	**Deep Cascade QA** Ming Yan [Yan et al. '18]	December 12th, 2018	52.01	54.64
4	**VNET** Baidu NLP [Wang et al. '18]	November 8th, 2018	51.63	54.37
5	**Masque NLGEN Style** NTT Media Intelligence Laboratories [Nishida et al. '19]	January 3rd, 2019	48.92	48.75
6	**BERT+ Multi-Pointer-Generator** Tongjun Li of the ColorfulClouds Tech and BUPT	December 31th, 2018	48.14	52.03
7	**SNET + CES2S** Bo Shao of SYSU University	July 24th, 2018	44.96	46.36
8	**Extraction-net** zlsh80826	October 20th, 2018	43.66	44.44
9	**SNET** JY Zhao	August 30th, 2018	43.59	46.29
10	**BIDAF+ELMo+SofterMax** Wang Changbao	November 16th, 2018	43.59	45.86

Tentative Timeline

- 2/13: Explore code and datasets. Finish proposal
- 2/20 - 2/27: Finish baseline models, get initial results on both dataset
- 3/6: Start to apply proposed ideas
- 3/27: Get an improved result
- 4/3: TBD

Thanks