Ronbun reading: A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

Danqi Chen and Jason Bolton and Christopher D. Manning

Saku Sugawara (U Tokyo, Aizawa-lab)

第 8 回最先端 NLP 勉強会

September 12, 2016

スライドは

- SpeakerDeck が見えづらい場合は
- http://penzant.net/files/snlp8-2016-09-12.pdf をごらんください

Ronbun

- A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
- Danqi Chen and Jason Bolton and Christopher D. Manning
- `https://arxiv.org/pdf/1606.02858v2.pdf`

  ACL2016 提出版から精度が向上している（ACL 発表はその内容）
- `http://cs.stanford.edu/people/danqi/`

  bib/paper/slide
- `https://github.com/danqi/rc-cnn-dailymail`

  only README???
- Figs are quoted from the original paper or the slides

Abstract

- CNN/Daily Mail 読解タスク (Hermann$^+$ 2015) のためのモデルとその分析
- ほぼ限界のスコアが出た（と主張している）

Background: Reading Comprehension

- 課題文とそれに関わる問いを読み、何らかの情報を返す
- 穴埋め、 課題文 $\rightarrow$ 選択肢 or 抜き出し などの形式

Passage (P) + Question (Q) $\longrightarrow$ Answer (A)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.....

Q What city is Alyssa in?

A Miami

Background: Reading Comprehension

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 + 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

Figure: 既存タスク例 (Rajpurkar$^+$ 2016)

Background: Big Data vs. Realistic

- 人手でデータを作ろうとするとどうしても小さくなる
  - MCTest (Richardson$^+$ 2013) [web]: 660*4 questions
  - ProcessBank (Berant$^+$ 2014) [web]: 585 questions
- 自動的に作るとたくさんデータができるものの、質が怪しい
  - CNN/Daily Mail (Hermann$^+$ 2015)
  - SQuAD (Rajpurkar$^+$ 2016) [web]
- 人手で比較的質の良いデータをたくさん作った例？
  - LAMBADA (Paperno$^+$ 2016) [web]
    まだちゃんと読んでないですがたぶんおすすめ

CNN/Daily Mail Dataset (DeepMind QA Dataset)

- Paper: Teaching Machines to Read and Comprehend `http://arxiv.org/pdf/1506.03340v3.pdf` (NIPS 2015)
- Site: `http://cs.nyu.edu/~kcho/DMQA/`

DeepMind Q&A Dataset

Hermann et al. (2015) created two awesome datasets using news articles for Q&A research. Each dataset contains many documents (90k and 197k each), and each doc average 4 questions approximately. Each question is a sentence with one missing word/phrase which can be found from the accompanying document/context.

The original authors kindly released the scripts and accompanying documentation to generate the datasets (see here). Unfortunately due to instability of WaybackMachi cumbersome to generate the datasets from scratch using the provided scripts. Furthermore, in certain parts of the world, it turned out to be far from being straight-forw WaybackMachine.

I am making the generated datasets available here. This will hopefully make the datasets used by a wider audience and lead to faster progress in Q&A research.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015).
Teaching machines to read and comprehend.
In Advances in Neural Information Processing Systems (pp. 1684-1692).

CNN

- Questions: here
- Stories: here

This dataset contains the documents and accompanying questions from the news articles of CNN. There are approximately 90k documents and 380k questions. I am making available 'questions/', which should be sufficient to reproduce the setting from the original paper, and 'stories/', which can be useful for other uses of this dataset.

Daily Mail

- Questions: here
- Stories: here

This dataset contains the documents and accompanying questions fr Daily Mail. There are approximately 197k documents and 879k questi available 'questions/', which should be sufficient to reproduce the set paper, and 'stories/', which can be useful for other uses of this datas

CNN/Daily Mail Dataset (DeepMind QA Dataset)

- CNN や Daily Mail の記事タイトルや見出しが該当箇所の要約になっている
- タイトルや見出しの entity 部分を穴にしてそれが何かを答えさせるタスク
- 記事はたくさんあるのでたくさん作れる

  (context, query, answer) で 1 単位
- 記事内容に答えが出てこないようなものは作らない

CNN/Daily Mail Dataset: Example

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

文脈文における correct answer の頻度

Top N	Cumulative %	
	CNN	Daily Mail
1	30.5	25.6
2	47.7	42.4
3	58.1	53.7
5	70.6	68.1
10	85.1	85.5

Table 2: Percentage of time that the correct answer is contained in the top N most frequent entities in a given document.

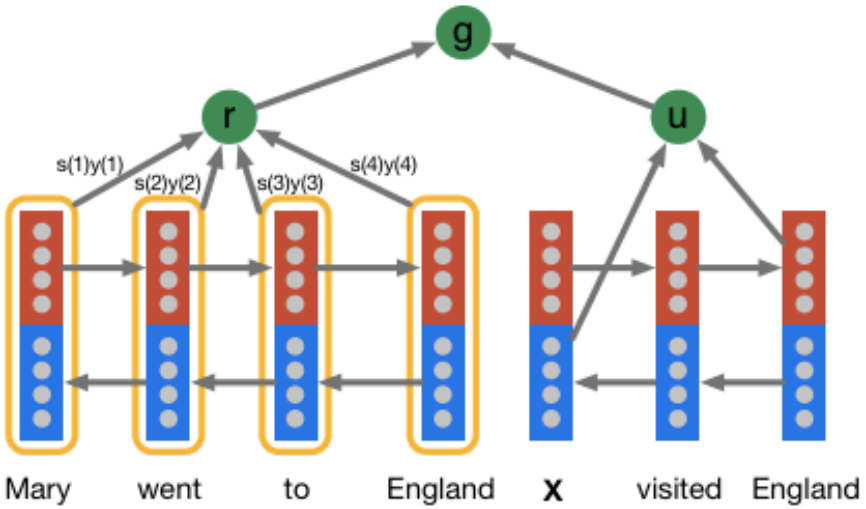
データセットの構成

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

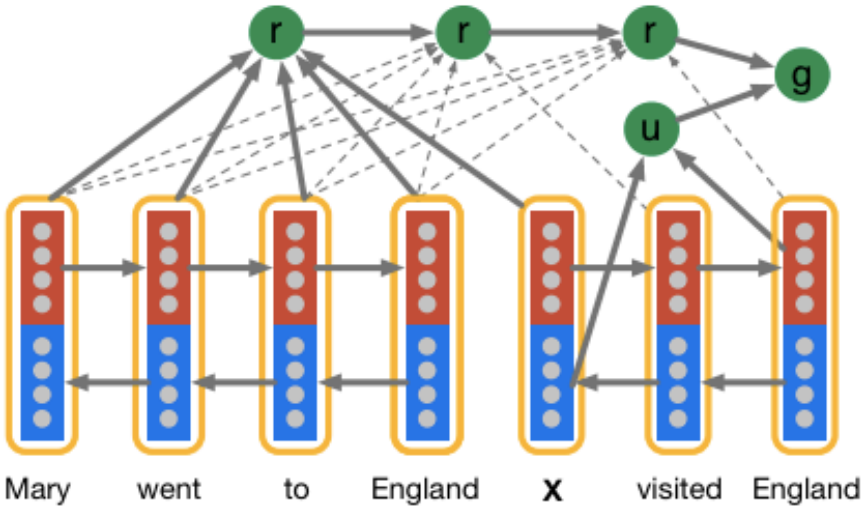
Table 1: Corpus statistics. Articles were collected starting in April 2007 for CNN and June 2010 for the Daily Mail, both until the end of April 2015. Validation data is from March, test data from April 2015. Articles of over 2000 tokens and queries whose answer entity did not appear in the context were filtered out.

たくさん

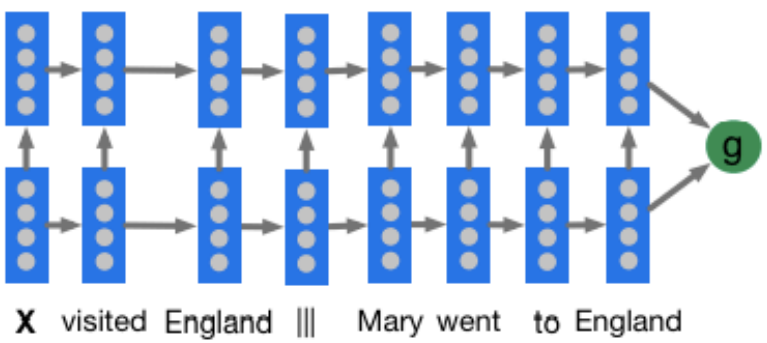
元論文で提案されている Neural Network Models



(a) Attentive Reader.



(b) Impatient Reader.



(c) A two layer Deep LSTM Reader with the question encoded before the document.

ここまで前座

本研究の貢献

- モデル：簡単なやつでだいたい良いスコアが出た
  1. Entity-Centric Classifier (比較・分析用？)
  2. End-to-end Neural Network (state of the art)
- 分析：だいたい良い分析をしてノイズを除いた上限を与えた
  - ランダムに選んだ 100 問を分類
  - エラーないし不明瞭なものが 25 問あった $\rightarrow$ 25% は無意味？

Entity-Centric Classifier

- 候補となる entities について以下を特徴にした vector f を構成
- 答えの entity の順位が高くなるように weight vector θ を学習

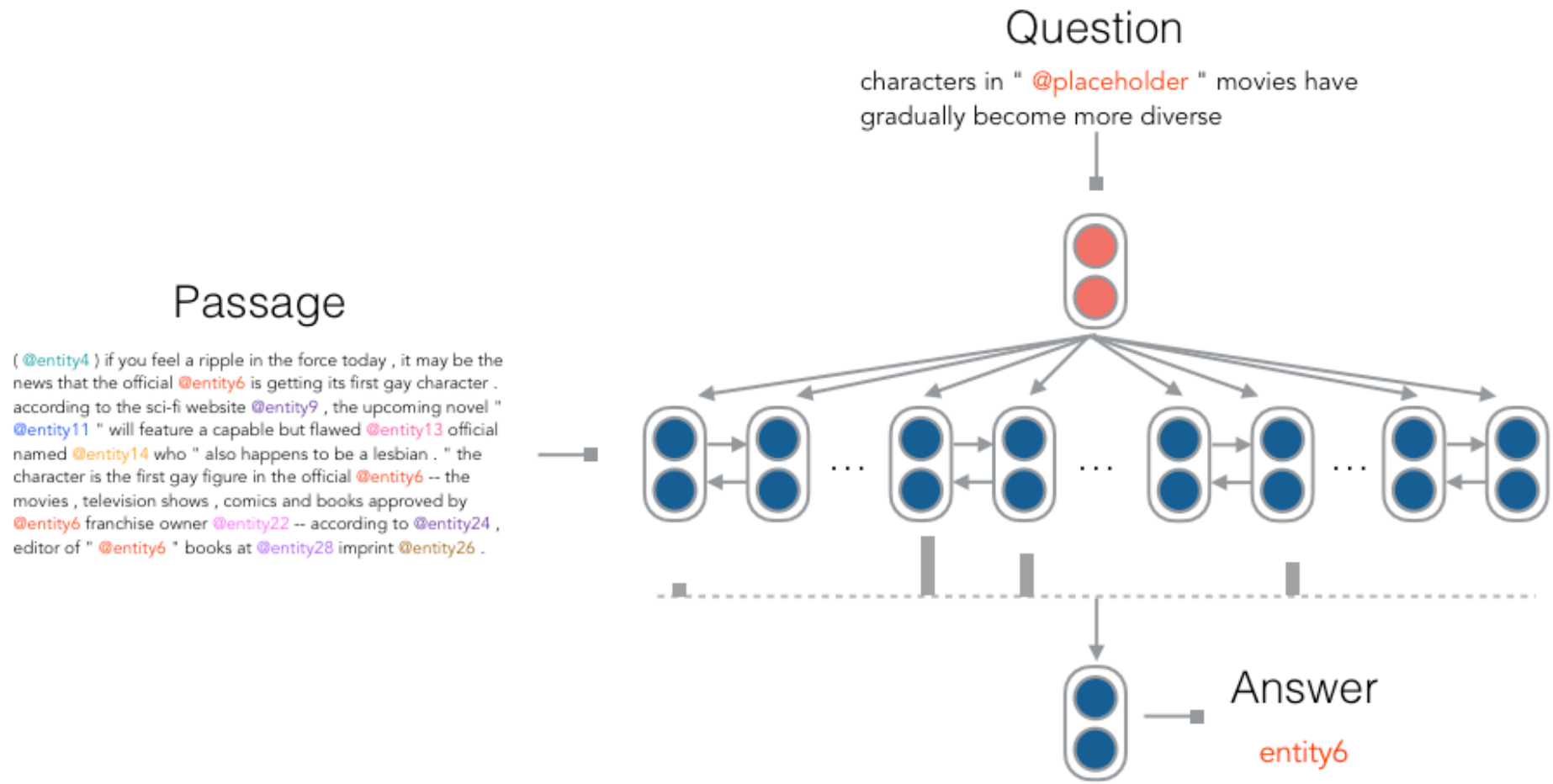
$$\theta^\top f_{p,q}(a) > \theta^\top f_{p,q}(e), \forall e \in E \backslash \{a\}$$

p: passage, q: question, e: entity, a: answer, E: entities
Algorithm: LambdaMart

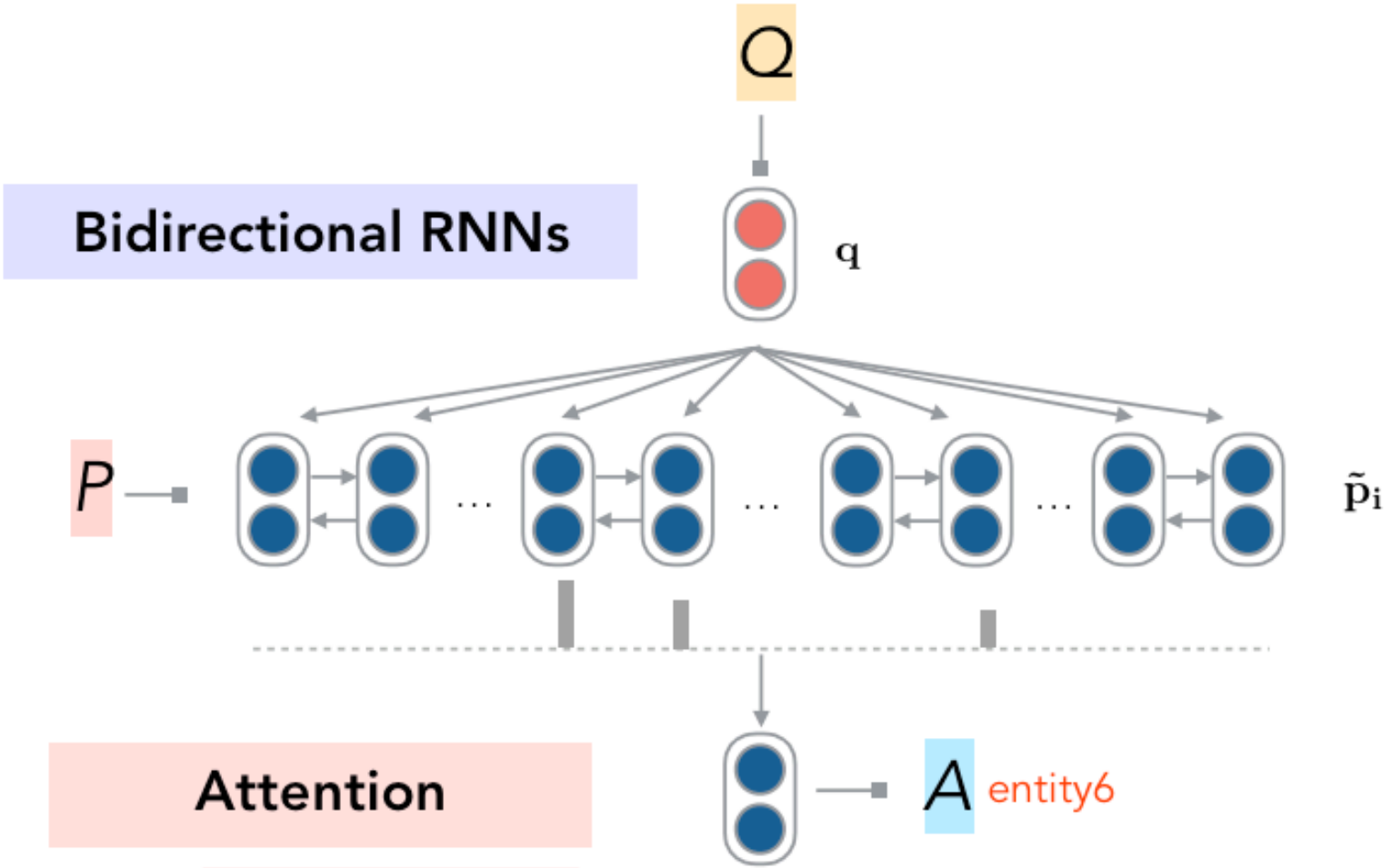
1. Whether *e* occurs in *P*
2. Whether *e* occurs in *Q*
3. Frequency of *e* in *P*
4. First position of *e* in *P*

5. *Whether e co-occurs with another Q word in P.*
6. word **distance**
7. **n-gram** exact match
8. **dependency parse** match

End-to-end Neural Network



End-to-end Neural Network



End-to-end Neural Network

1. Encoding
   - Bi-directional RNN + GRU
     ACL 版では LSTM だったが改訂版 (v2) では RNN+GRU に変更
   - 1. $\mathbf{h}_i^{\mathrm{R}} = \mathrm{RNN}(\mathbf{h}_{i-1}^{\mathrm{R}}, \mathbf{w}(\mathbf{p}_i))$, $\mathbf{h}_i^{\mathrm{L}} = \mathrm{RNN}(\mathbf{h}_{i+1}^{\mathrm{L}}, \mathbf{w}(\mathbf{p}_i))$
   - 2. $\mathbf{p}_i = \mathrm{concat}(\mathbf{h}_i^{\mathrm{R}}, \mathbf{h}_i^{\mathrm{L}})$
2. Attention
   - 1. $\alpha_i = \mathrm{softmax}_i\ \mathbf{q}^\top \mathbf{W}_s \mathbf{p}_i$
   - 2. $\mathbf{o} = \sum_i \alpha_i \mathbf{p}_i$
     α: prob. distribution (=attention), $\mathbf{q}$: question embedding, $\mathbf{p}_i$: contextual embedding for p_i (i-th word in the passage), $\mathbf{W}_s$: weight matrix used for a bilinear term (it frexibly computes a similarity between $\mathbf{q}$ and $\mathbf{p}_i$), $\mathbf{o}$: output vector
3. Prediction
   - $a = \mathrm{argmax}_{a \in p} W_a^\top \mathbf{o}$

Diffs from previous model (Hermann$^+$ 2015)

1. bilinear term using $\mathbf{W}_s$, instead of a $\tanh$ layer
   - similarity between $\mathbf{q}$ and $\mathbf{p}_i$ の表現の仕方を変えた
     柔軟性が上がった？
2. $\mathbf{o}$: output vector を最終的な予測に使う過程で余計な計算を挟まないようにした
   - 元のモデルでは変なレイヤーをいろいろ噛ませていた
3. prediction 対象の vocaburaly を entity だけにした
   - 元のモデルは出現するすべての語を候補にしていた

Result

Model	CNN		Daily Mail	
	Dev	Test	Dev	Test
Frame-semantic model †	36.3	40.2	35.5	35.5
Word distance model †	50.5	50.9	56.4	55.5
Deep LSTM Reader †	55.0	57.0	63.3	62.2
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (window memory) ‡	58.0	60.6	N/A	N/A
MemNNs (window memory + self-sup.) ‡	63.4	66.8	N/A	N/A
MemNNs (ensemble) ‡	66.2*	69.4*	N/A	N/A
Ours: Classifier	67.1	67.9	69.1	68.3
Ours: Neural net	72.5	72.7	76.9	76.0
Ours: Neural net (ensemble)	76.2*	76.5*	79.5*	78.7*
Ours: Neural net (relabeling)	**73.8**	**73.6**	**77.6**	**76.6**
Ours: Neural net (relabeling, ensemble)	**77.2***	**77.6***	**80.2***	**79.2***

Analysis - classifier model - ablation

Features	Accuracy
Full model	67.1
– whether e is in the passage	67.1
– whether e is in the question	67.0
– frequency of e	**63.7**
– position of e	65.9
– n-gram match	**60.5**
– word distance	65.4
– sentence co-occurrence	66.0
– dependency parse match	65.6

Analysis - sampled questions

Category	Question	Passage
Exact Match	*it 's clear @entity0 is leaning toward* @**placeholder** , says an expert who monitors @entity0	. . . @entity116 , who follows @entity0 's operations and propaganda closely , recently told @entity3 , *it 's clear @entity0 is leaning toward* @**entity60** in terms of doctrine , ideology and an emphasis on holding territory after operations . . . .
Paraphrase	@**placeholder** says he understands why @entity0 wo n't play at his tournament	. . . @entity0 called me personally to let me know that he would n't be playing here at @entity23 , " @**entity3** said on his @entity21 event 's website . . . .
Partial clue	a tv movie based on @entity2 's book @**placeholder** casts a @entity76 actor as @entity5	. . . to @entity12 @entity2 professed that his @**entity11** is not a religious book . . . .

Analysis - sampled questions

Multiple sent.	he 's doing a his - and - her duet all by himself , @entity6 said of **@placeholder**	. . . we got some groundbreaking performances , here too , tonight , @entity6 said . we got **@entity17** , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . . . .
Coref. Error	rapper **@placeholder** " disgusted , " cancels upcoming show for @entity280	. . . with hip - hop star **@entity246** saying on @entity247 that he was canceling an upcoming show for the @entity249 . . . . (but @entity249 = @entity280 = SAEs)
Hard	pilot error and snow were reasons stated for **@placeholder** plane crash	. . . a small aircraft carrying **@entity5** , @entity6 and @entity7 the @entity12 @entity3 crashed a few miles from @entity9 , near @entity10 , @entity11 . . . .

Analysis - result

Exact match

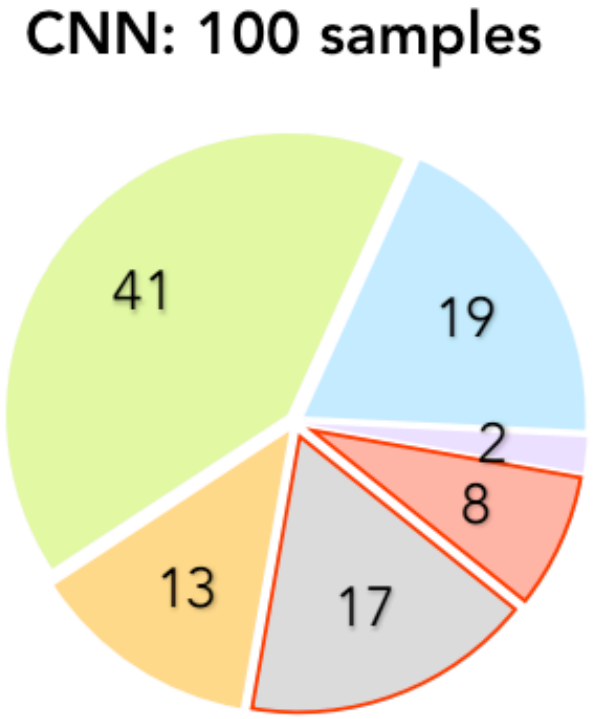
Paraphrasing

Partial clue

Multiple sentences

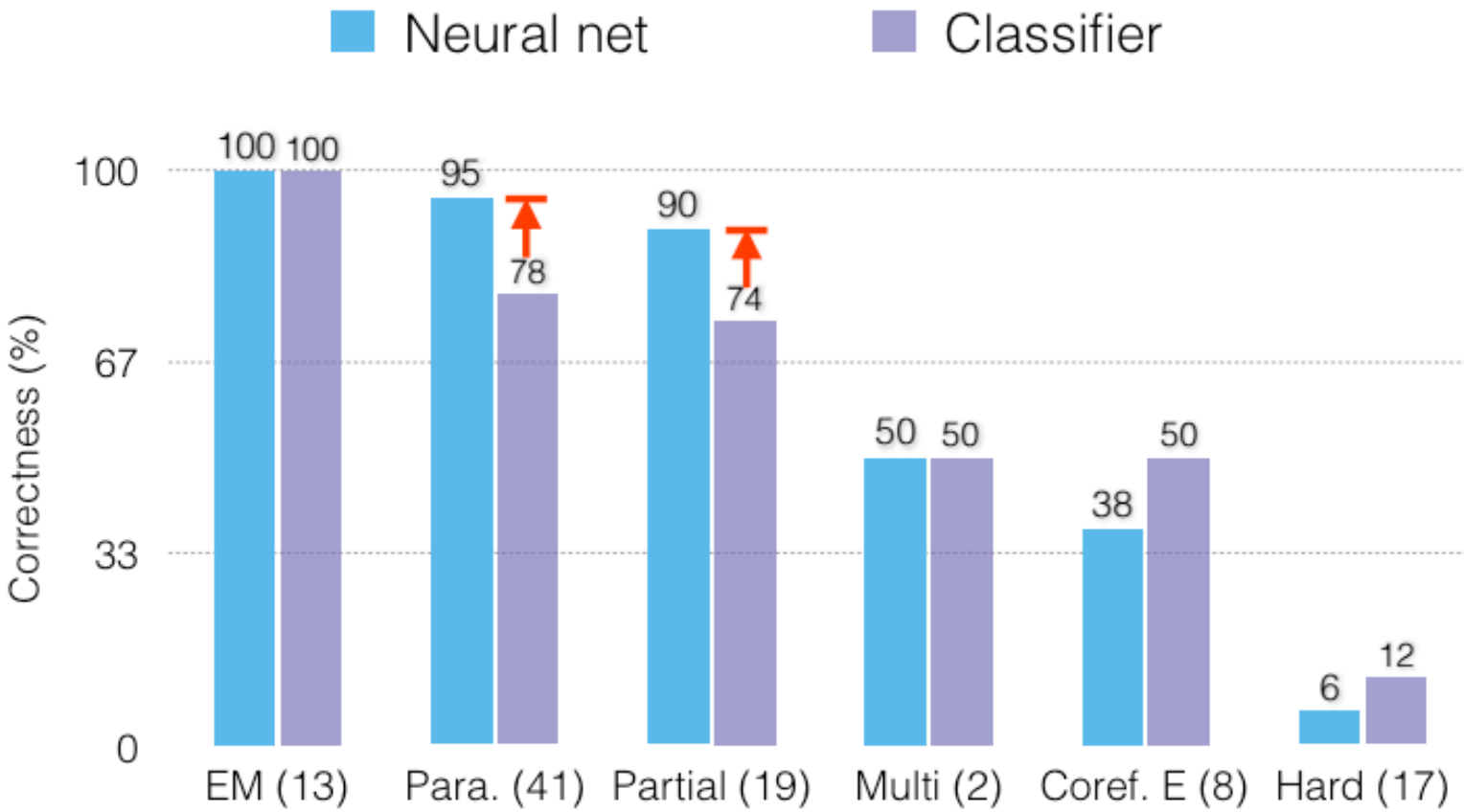
Coreference errors

Ambiguous / hard



neural net	**73.8**	73.6
neural net (ensemble)	**77.2**	**77.6**

Analysis - accuracies for each category



まとめ

- シンプルなモデルが良かった

  semantic matching を学習するには neural model がやっぱり良い（単なる classifier と比べると）
- CNN/Daily Mail はほとんど頭打ち: データにノイズが多い、自動で作れたのは良いけど質も大事（本当に読解的な推論を測るのに役立つのか？）
- こういうデータセットを否定する必要はなく、より realistic なデータセットのための学習データとして活かせるはず
- いろいろデータセットが増えてるし reading comprehension 流行してますね
- 感想: 実データ見て丁寧に分析するのが大事