Extractive NarrativeQA with Heuristic Pre-Training

Lea Frermann
Amazon
uedi@frermann.de

Abstract

Although advances in neural architectures for NLP problems as well as unsupervised pre-training have led to substantial improvements on question answering and natural language inference, understanding of and reasoning over long texts still poses a substantial challenge. Here, we consider the task of question answering from full narratives (e.g., books or movie scripts), or their summaries, tackling the NarrativeQA dataset (NQA; Kocisky et al. (2018)). We introduce a heuristic *extractive* version of the data set, which (a) leads to a large data set of questions with answers *extracted* from the summaries; (b) allows us to tackle the more feasible problem of answer extraction (rather than generation). We train systems for passage retrieval as well as answer span prediction using this data set, on top of pre-trained BERT embeddings. We show that our setup leads to state of the art performance on summary-level QA. On QA from raw narrative text, we show that our model performs comparatively to previous models. We analyze the relative contributions of pre-trained embeddings and the extractive training paradigm, and provide a detailed error analysis.

1 Introduction

With recent advances in machine learning techniques, the availability of sizable data sets as well as compute power, natural language processing has made impressive advances across a variety of NLP tasks. A striking gap between machine and human performance, however, remains the ability to *comprehend* text and make *inferences* over multiple pieces of information.

Automatic question answering (QA) from text has received much recent attention as a task designed towards bridging this gap. A variety of question answering tasks and data sets with different levels of difficulty have been proposed recently, ranging from questions paired with short, relevant documents containing immediately inferable answers (SQUAD; Rajpurkar et al. (2016)), over questions to be answered from sets of documents and requiring to connect facts through multi-step inferences (WikiHop; Welbl et al. (2018)) to naturally occurring questions as Google search queries, paired with sets of Wikipedia pages (Natural Questions; Kwiatkowski et al. (2019)).

Common characteristics of all these data sets are (1) a large number of training examples in the order of tens- to hundreds of thousands training and test examples; (2) *extracted* answers which can be pin-pointed in the reference documents; (3) the reference documents from which answers are derived are of comparatively short length (e.g., an average of 100 tokens per reference for WikiHop, vs 60K tokens in NQA). All recently proposed successful QA systems were trained in a supervised way, heavily relying on the availability of answer-annotated data sets as described above.

In this work we consider the highly challenging task of narrative question answering (NQA), as introduced by Kocisky et al. (2018). In NQA, a system is presented with a question on the plot of a narrative (a book or a movie) and produces a free-text answer given the raw book or movie script text. The data set was created by pairing each original narrative with a human-created summary. Based on the summaries, a large set of question-answer pairs was created through crowd sourcing. The questions were derived from the summaries to deliberately avoid answers to be straightforwardly extractable from the full narrative texts.

Several interesting challenges arise in NQA: (1) the reference documents are orders of magnitude longer than in previous tasks and demand long-range reasoning and inference; (2) answers are *abstractive* and as such not necessarily ver-

batim in the reference documents; (3) although answers are typically localized in the summary, the corresponding answer in the book often requires reasoning across paragraphs or even chapters; (4) the size of the data set, shown in Table 2, is comparatively small making supervised training challenging.

This paper explores the utility of heuristic, but inexpensive training data sets for NQA. We formulate NQA as an *extractive* question answering tasks, leveraging the fact that due to the generation of the data set answers tend to be extractable locally from the summary text (cf., Table 1 for examples). While ultimately an abstractive system, which synthesizes an answer based on information in the text, is desirable, a conceptually simpler extractive approach can serve a first and more feasible step towards the goal of answer generation. Our evaluation shows that our extractive system performs competitively on summary- and book-level NQA.

The main contribution of this paper is a heuristic extractive NQA data set created by leveraging characteristics of the generative process of the original NQA dataset. Specifically, since question-answer pairs were synthesized based on the summaries, we hypothesize that typically a single summary sentence (or subspan thereof) contains the answer to a question. We develop heuristics to retrieve those spans.

Based on our heuristic extractive data set we train models for two tasks: (1) Question-based sentence retrieval, which, given a question, selects relevant book passages for a question (which may serve as input to a sophisticated QA model); and (2) SQUAD-style answer extraction, where the system learns to point to the beginning and end of the answer in the reference text. We train systems for sentence-retrieval and answer selection on top of pre-trained BERT embeddings (Devlin et al., 2018).

We train question answering system on summary-question-answer tuples, and evaluate the systems both on summary references as well as on the full book text. While a variety of systems have been proposed for summary-level based NQA, to the best of our knowledge, we are the first to tackle the full NQA challenge where questions are answered from the complete, raw narrative (book or movie script) text. Our system achieves state-of-the-art results on summary-level answer extraction, and on the book-level outperforms previously proposed span extraction model, and performs competitively previous neural reasoning models.

In summary, our contributions are:

1. Augmentation of existing (sparse) data sets with heuristic, inexpensive and supervised training data, with an application to extractive question answering for NQA

2. The first reported results on the full narrative-based question answering challenge since the introduction of the data set (Kocisky et al., 2018)

3. State-of-the-art results on the summary level NarrativeQA benchmark

4. An analysis of common errors shedding light on shortcomings in model performance as well as evaluation

2 Task Description

We first describe the original NarrativeQA data set and task. (Kocisky et al., 2018) provides a testbed for question answering on raw narrative text. It consists of over 1,567 publicly available full-length narrative documents (books or movie scripts), each paired with a human-created plot summary. For each document a set of question-answer pairs was collected by presenting human annotators with the summary. The annotators generated a set of questions together with free-text answers. For each document, 30 questions were collected, each with two free-text answers (provided by distinct annotators).

The data set consists of a total of 46,765 question-answer pairs. Considering the variety in question types, narrative styles (books and movie scripts of different genres), sheer length of the documents, and the fact that answers need to be synthesized, this data set is too small to train an NQA model in an unsupervised way.

We alleviate the above issues in two ways. First, we incorporate prior knowledge in the form of pre-trained word embeddings (Devlin et al., 2018). Second, we recognize that by construction of the data set, answers to questions can generally be localized in the summaries, even though the free-text answers are typically not found verbatim in the summary. We leverage this property to con-

Question	why does nora track mark down
References	R1: malcom ' s suicide ; R2: to confront him after malcolm commits suicide
Extracted	Nobody knows the true identity of hard harry or happy harry hard-on , as mark refers to himself , until nora diniro (mathis) , a fellow student , tracks him down and confronts him the day after a student named **malcolm commits suicide** after harry attempts to reason with him .
Question	why did the couple visit medium shaun san dena in pasadena in 1969
References	R1: their son has been hearing voices from evil spirits; R2: because their son was hearing evil spirits voices
Extracted	in 1969 pasadena , california , a couple seeks the aid of the medium shaun san dena (flor de maria chahua) saying their **son (shiloh selassie) has been hearing evil spirits ' voices** after stealing a silver necklace from gypsies .
Question	how was hadley s hope colony destroyed
References	R1: the nuclear blast from the damaged power plant; R2: an explosion
Extracted	all four escape moments before the station explodes with the colony consumed by the **nuclear blast** .

Table 1: Example questions from the NarrativeQA data set, with human provided free-text reference answers (Gold), as well as the most relevant automatically extracted sentence-level answer from the summary (Extracted) and the most relevant sub-sentence level span (boldface).

struct *extractive* data sets for sentence-level and sub-sentence level answer extraction.

3 Extractive NarrativeQA

We derive data sets for supervised query-based sentence retrieval (Section 3.1), and answer span extraction (Section 3.2).

3.1 Sentence Retrieval Data Set

For each question, and its corresponding summary, we proceed as follows. We first obtain a relevance score of each summary sentence s to the input question q, as follows. We concatenate the question[1] q with both human-created free text answers $a1, a2$,

$$z = [q; a1; a2], \tag{1}$$

and obtain a relevance score of each summary sentence s w.r.t. z by passing both through Google's Universal Sentence Encoder (USE) (Cer et al., 2018) and computing the cosine similarity between the encodings,

$$rel_z(s) = cos(use(z), use(s)). \tag{2}$$

We can thus obtain a ranking of summary sentences w.r.t., their relevance to input qa-pair z. Such a ranking can serve as a sentence or passage retrieval system, providing input to a more sophisticated question answering model . We further use these relevance scores as a basis for heuristic answer-span annotation as described in the following section. Example questions, together with the most relevant retrieved sentence, are shown in Table 1.

[1]We remove the question mark and the first word if it indicates a wh-question

3.2 Answer Span Prediction Data Set

Although sentence retrieval is an important step towards question answering from narratives, ultimately more flexibility in granularity of the retrieved information is necessary. To this end, building on sentence-level relevance scores, we create a dataset of answer-annotated summaries, selecting as answer the most relevant word sequence to a question q in the summary. We extract an answer span using the following back-off strategy:

1. if available, return an exact match of one of the reference answers
2. if unsuccessful: considering the three most question-relevant sentences as determined by the USE (Section 3.1) find the longest substring bounded by content words in the query
3. if unsuccessful: considering any sentence in the summary, return the longest substring bounded by content words in the query

	train	valid	test
# QA-pairs	32,170	3,461	10,557
# documents	1,102	115	355

Table 2: Statistics of the NarrativeQA data set (Kocisky et al., 2018). We obtain a heuristic answer match for each original question, and maintain the original train/valid/test splits.

Given a dataset of questions, paired with documents containing the answers, we are able to train SQUAD-style answer prediction systems (cf., Section 5). Examples of retrieved answer spans are shown in bold face in Figure 1.

4 Experiment Setup

For our experiments, we *train* systems for sentence retrieval, and answer span prediction on questions paired with summaries, and answers as heuristically extracted text passages from the summaries as described in Sections 3.1 and 3.2. We *test* sentence retrieval and answer span prediction experiments on both summary level data, and full narrative texts. We evaluate our extractive model predictions against the original, *abstractive* NarrativeQA gold answers using the exact same evaluation setup as all the previous work we compare against.

Our experiments investigate (a) the effectiveness of a heuristic training data set on sentence retrieval and answer span prediction in the context of NQA; (b) the extent of generalization of systems trained on summary data to book full texts; and (c) the utility of strong prior knowledge in the form of pre-trained word embeddings. We train sentence retrieval and span prediction models on top of pre-trained BERT embeddings (Devlin et al., 2018). We start by explaining the model architecture, before we describe experiments and results on the extractive NQA tasks.

4.1 BERT

BERT embeddings (Devlin et al., 2018) are word representations which have been pre-trained on an enormous training corpus on unsupervised word- and sentence prediction tasks. They have been shown to encode substantial semantic and syntactic information, and have been efficiently fine-tuned towards a variety of NLP tasks leading to new state-of-the-art results (Devlin et al., 2018).

Here, we fine-tune BERT embeddings for NQA sentence retrieval and answer span selection, as described in the following sections.

	accuracy	precision	recall	f1
$p_{rel} > 0.5$	0.87	0.88	0.83	0.86

Table 3: Results on summary-level sentence-relevance classification on the NQA test set of 25K question-answer pairs. We set the relevance threshold to $p > 0.5$.

5 Sentence Retrieval

Given a query and a reference text, the task is to retrieve the most relevant sentences from the reference to the query. For each sentence in the reference text, we individually compute a relevance probability.

Approach Given a large set of sentence-question pairs, we train a relevance prediction model on top of BERT embeddings. Following closely the architecture for BERT-based sentence classification, our system takes as input the BERT-embedded query q concatenated with a single BERT-embedded summary sentence s,

$$z = [START]enc(q)[SEP]enc(s), \quad (3)$$

which is passed through a single linear layer followed by a softmax layer.

The for each sentence-query pair we obtain a relevance score $\in [0, 1]$, from which we can derive a summary sentence ranking by query relevance, and from this ranking retrieve the top n most relevant sentences for further predictions.

We use all default parameters from the original BERT implementation.[2]

Summary-level results We apply our model to the book summaries from test data set of NarrativeQA. We evaluate the extent to which truly relevant sentences (as extracted by our heuristic method) were assigned a relevance probability $p > 0.5$. Results are shown in Table 3, and show that the model is very well able to detect the most relevant summary sentence for a question.

Book-level results We also apply the model to the considerably harder task of NQA on full narrative documents, computing a question-specific relevance score for each sentence in the document.

[2] https://github.com/google-research/bert

	p@1	p@5	MRR
BM25f	10.53	51.42	0.276
BERT	13.80	53.02	0.305

Table 4: Fraction of correct answers contained in the top {1 / 5} answer candidates, and MRR of the correct answer in passages retrieved by the BERT-based retrieval method (BERT) or an IR method (BM25f).

Note that we cannot evaluate retrieval scores directly, because we do not have access to reference sentence-relevance scores.

Instead, we treat our system as a passage retrieval model given an input question. As an approximation to the quality of the retrieved passages we compute the extent to which the correct answer is found in the N most frequent answer candidates.[3]

We compare our BERT-retrieval with an IR-style retrieval system (BM25f; Zaragoza et al. (2004)) which retrieves text passages of five consecutive sentences based on word token and character mention overlap with the question. From both systems, we retrieve the 20 most relevant sentences, each in a context of ± 2 sentences.

The results are shown in Table 4. We can observe that the memory network with BERT-based retrieval outperforms the IR retrieval-based model. We will also incorporate this model as a passage-preselection module for book-level answer span prediction in Section 6.

Qualitatively, we observed that most book sentences receive a very low relevance probability, which makes the model amenable for the task of narrowing down the context to few relevant passages. E.g., on average across all books, only 1.4% of all sentences are predicted as relevant with $p >= 0.8$ and 4.3% with $p >= 0.01\%$.

6 Answer Span Prediction

Given a question and a reference text (summary or full narrative), the task is to predict a continuous sub span of the reference text as the answer to the question.

Approach We fine-tune BERT embeddings for answer extraction, similar to the approach for BERT-based SQUAD question answering in Kocisky et al. (2018). Given a query q and a text passage c, we map both to BERT embeddings, and concatenate the embedded representations,

$$z = [START]enc(q)[SEP]enc(c). \quad (4)$$

We train two pointers, which, given $enc(q)$, select the start and end word of the answer span in $enc(c)$, respectively. Pointing to the $[START]$ token, the model also has the capacity to predict no answer at all.

While we use the whole summaries as contexts for summary-based QA, considering full narrative texts is prohibitive. To this end, we leverage the sentence retrieval model from Section 5 to obtain a subset of relevant sentences. In our experiment we retrieve the 100 most likely sentences given a question, each in a context of ± 2 sentences, resulting in contexts of (up to) 500 sentences per question.

Even after this pre-selection, memory constraints prohibit processing of the full contexts, or summary texts. Following prior work (Kocisky et al., 2018), we limit context length to a maximum of 384 words, split the original reference documents into multiple segments, and pass each segment individually as context, and return the most likely span across all passages as an answer. For each test input, we return the most likely non-empty answer returned by the model.

In order to disentangle the contribution of powerful BERT embeddings from the utility of our heuristic training corpus, we also trained an answer extraction model using SQUAD-V2.0 training data (Rajpurkar et al., 2018) (BERT SQUAD). We train the models using either the full SQUAD data set, or a random subset of 31,000 training items, comparable in size to our heuristic training data set. On the one hand, this data set is a gold-standard of prefect context-span to answer correspondences. On the other hand, the data is out-of-domain and thus potential less informative for a NQA task.

We evaluate the predicted answers against the human-provided free-text answers using BLEU and METEOR scores. We report results using summaries as contexts for QA, and on using full narratives as contexts, and compare against recent competitive models.

Summary-level Results Table 5 displays summary-level answer span extraction results for a set of previously proposed models (top), the BERT-based span prediction model trained on

[3] We evaluate our system only in the context of *who?* questions with an entity as answer and consider all book entities as candidate answers.

model	BLEU-1	BLEU-4	METEOR	Rouge-L
BiDAF Span Prediction (Kocisky et al., 2018) (test)	33.45	15.69	15.68	36.74
DecaProp (Tay et al., 2018) (test)	42.00	23.42	21.80	44.69
ConZNet (Indurthi et al., 2018) (test)	42.76	22.49	19.24	46.67
BERT SQUAD train (valid)	36.18	16.42	24.15	50.13
BERT SQUAD train (test)	36.22	17.14	23.61	48.58
BERT SQUAD train 31K (valid)	42.69	21.12	20.78	48.00
BERT SQUAD train 31K (test)	40.71	20.60	19.78	45.06
BERT heur (valid) [2019-05-10]	50.04	23.25	26.30	58.18
BERT heur (test) [2019-05-10]	**50.36**	**24.24**	**27.09**	**58.50**

Table 5: Summary-level answer extraction results by previous models and our systems (BERT) trained on out-of-domain SQUAD data, and our heuristic data set.

model	BLEU-1	BLEU-4	METEOR	Rouge-L
BiDAF Span Prediction (Kocisky et al., 2018) (test)	5.68	0.25	3.72	6.22
Att sum 10 chunks (Kocisky et al., 2018)(test)	**19.09**	1.81	4.29	14.03
Att sum 20 chunks (Kocisky et al., 2018)(test)	19.06	**2.11**	4.37	14.02
BERT SQUAD train (valid) **shuf**	9.16	0.93	4.28	11.09
BERT SQUAD train 31K (valid) **shuf**	8.87	0.97	3.23	9.96
BERT SQUAD train (test)	9.06	1.03	4.29	10.58
BERT SQUAD train 31K (test)	9.23	1.47	3.55	10.29
BERT heur **shuf** (valid) [2019-05-17] 100-cxt2 FIN	11.91	1.75	4.79	13.98
BERT heur **shuf** (test) [2019-05-17] 100-cxt2 FIN	11.63	1.69	4.93	14.38
BERT heur (valid) [2019-05-28]	12.07	1.70	5.00	14.83
BERT heur (test) [2019-05-28]	12.26	2.06	**5.28**	**15.15**

Table 6: Book-level answer extraction results by previous models and our systems (BERT) trained on out-of-domain SQUAD data, and our heuristic data set.

SQUAD data (center), and the same model trained on our heuristic extractive NQA corpus (bottom).

BiDAF is a span prediction model, conceptually similar to our own and was proposed as a baseline method in (Kocisky et al., 2018). DecaProp (Tay et al., 2018) is a neural network which, through particularly dense connections between neighboring layers, is designed to distill information from hierarchical passage representations (over words, sentences, and paragraphs). CoZNet (Indurthi et al., 2018) is a neural network architecture designed to 'zoom into' relevant passages of contiguous, long text passages, using co-attention on query and passage and reinforcement learning with answer generation as target. The latter models *generate*, rather than extract, an answer. All models were evaluated against the human free-text answers.

We observe that the model trained on the heuristic data set substantially outperforms all prior work. The model trained on SQUAD data compares poorly against all other models, demonstrating that the prior information from BERT embeddings by themselves do not automatically lead to improvements on NQA. Interestingly, the SQUAD-data trained model perform better with fewer data (31K) compared with the full training data set, suggesting that fitting the model to SQUAD-data prediction decreases its generalization ability to out-of-domain NQA test data. The positive results with the heuristic training corpus suggests that a heuristic and potentially noisy in-domain data set is of great utility for summary-level answer span extraction.

Book-level Results Although a range of prior models have been proposed for summary-level QA, to the best of our knowledge, no results on full document-level NQA have been published since the original benchmark paper. We compare against the most relevant, and most competitive system described in the original paper (Kocisky et al., 2018). All results are shown in Table 6.

	Question	who is mark hunter
1 ✓	Gold	he is a high school student in phoenix
	Model	high school student
	Question	why do more students tune into mark s show
2 ✓	Gold	mark talks about what goes on at school and in the community
	Model	speaks his mind
	Question	why do faulkland and julia always fight
3 ✓	Gold	he thinks she s unfaithful
	Model	jealous suspicion . he is constantly fretting himself about her fidelity
	Question	who was murphy s ghost
4 ✗	Gold	cooper from the future
	Model	a poltergeist
	Question	what name does this extended meditation focus on
5 ✗	Gold	z. marcas
	Model	the nature of human names

Figure 1: Example questions and top-ranking model-extracted answer from the summaries. Top: examples the model answered correctly (✓); Bottom: questioned answered incorrectly by the model (✗).

We compare our own model trained on the heuristic training corpus (bottom), against another span prediction model, Bi-Directional Attention Flow (BiDAF; Seo et al. (2016)), as reported in Kocisky et al. (2018), as well as their most competitive model, an adaptation of the Attention Sum Reader (Kadlec et al., 2016) (AS Reader). AS Reader follows an encoder-decoder architecture with attention, where the decoder is an LSTM sequence decoder which can synthesize an answer (rather than extract). Both prior models are combined with a passage pre-selection method (similar to our own), which is based on tf-idf based cosine similarity of answers (for training sets) and questions for (test sets). Like for the book-level task, we also compare our architecture fine-tuned on out-of-domain, but high-quality, SQUAD QA training data.

First, we can observe that our model outperforms the conceptually similar span extraction model. The AS Reader performs similarly to our model, with the ranking depending on the metric used. Our model outperforms all previous metrics in terms of METEOR score. METEOR includes synonym matching and as such recognizes semantically similar predictions to the gold standard. The error analysis (Section 7), provides a variety of examples which demonstrate that model predictions are indeed often correct, despite having little word overlap with the gold standard. Like in the summary-level evaluation, models trained on our own corpus outperform the SQUAD-based models, suggesting again the utility of training on easily obtainable, inexpensive but heuristic in-domain data.

7 Error Analysis

We inspect a variety of examples on both summary- and narrative level QA to shed lights on shortcomings of the model and evaluation.

Figures 1 displays example questions with gold and model predicted answers from the summaries as reference documents. Example 1 shows a case where the correct answer is conceptually simple and easily extractable. Examples 2 and 3 contain answers which are complex concepts as indicated by the more verbose human and model-produced answers. Still, the model predictions are correct in both cases. Examples 4 and 5 shows examples with incorrect model answers. Although predicted span is clearly semantically related to the question, it is not a valid response.

We show questions with human and model predictions based on passages from the full narrative in Figure 2. We also include the local context from which the model answer was extracted (the full context is up to 500 sentences long). Examples 1, 2, 6 and 7 are predicted correctly by the model. Note that the predicted answer for examples 1, 6 and 7 have no (or little) lexical overlap with the gold answer, although the prediction is clearly correct as supported by the context. Example 3 illustrates a case where the model-predicted answer is wrong, however, a the proposed passage

1 ✓	Question	what is tom trying to desperately get working
	Gold	his latest invention
	Model	photo telephone
	Context	[...] this took him the better part of a week , but he did not regret the time , for while he was working he was busy making plans for **his newest invention – the photo telephone** . [...]
2 ✓	Question	what is dubuches passion besides painting
	Gold	music
	Model	music
	Context	[...] his landscapes were at least conscientiously painted , excellent in intention ; **but his real passion was music** , a madness for music , a cerebral bonfire which set him on a level with the wildest of the band . [...]
3 ✗	Question	how does claude die
	Gold	he hangs himself
	Model	knocked down by a little hand-truck
	Context	[...] **claude** , who was stepping back , **narrowly escaped being knocked down by a little hand-truck** which two big full-bearded fellows brought up [...]
4 ✗	Question	how does mr. peter s feel about tom not allowing him access to the patents
	Gold	he gets angry
	Model	embarrassed
	Context	[...] ” who are these men ? ” **tom felt a little embarrassed** over what he had said . ” could n’t he make him pay ? ” asked tom . [...]
5 **?**	Question	what were owen and george found doing the morning after brenda died
	Gold	they were laughing and singing
	Model	singing snatches of songs
	Context	[...] george and owen were singing snatches of songs and gibbering like two great apes .
6 ✓	Question	how long does jimmy gator have to live
	Gold	just a few months
	Model	two months
	Context	[...] [[jimmy gator]]: i have about **two months** , i have no time . [...]
7 ✓	Question	what happened to the doctor who was called to the house after mortimer s brothers went insane
	Gold	he collapsed at the sight of brenda s body
	Model	fell into a chair
	Context	[...] you will recollect that mortimer tregennis , in describing the episode of his last visit to his brother ’s house , remarked that **the doctor on entering the room fell into a chair ?** [...]
8 ✗	Question	how does linda try to commit suicide
	Gold	by overdosing on earl s prescription medication
	Model	slams the door
	Context	[...] your boyfriend let me in , i just knocked on the door – claudia he ’s not my boyfriend . linda ’s merceds - moments later **she slams the door** . earl .... getthat on the tv [...]
8 ✗	Question	what is ironic about donnie s teeth being knocked out when he falls from the pole
	Gold	he no longer has to worry about getting braces on his teeth
	Model	donnie ’s mouth is full of blood and his teeth
	Context	[...] . he.he . **donnie ’s mouth is full of blood and his teeth are broken** ... [[donnie]]: my teeff ... my teeef .... [[jim kurring]]: you ’re ok [...]

Figure 2: NQA from whole narrative texts: example questions, gold human-created answers, and model predictions together with local context from which the predictions were extracted. We indicate whether an answer is correct (✓), incorrect (✗) or undecidable (**?**).

covers a situation which is similar to the correct answer (nearly escaping a potentially deadly situation, rather than real death of the same person). Example 4 is a wrong prediction, a result of confusing semantic roles of the participants. Example 5 seems to be correct, however, from the context it is not clear whether the extracted passage indeed refers to *the morning after brenda died*. Examples 8 and 9 show wrong predictions, but the extracted contexts are still meaningful and semantically relevant to the query.

Overall, the error analysis suggests (a) that purely data-driven models tend to overly rely on surface semantic similarity. It furthermore suggests, however, that the automatic evaluation scores BLEU, METEOR and Rouge, which rely on word overlap, are overly conservative regarding the output of our model. A series of recent papers discussed problems of comparing models on abstractive NLI tasks using automatic metrics as the ones listed above (Novikova et al., 2017; Chaganty et al., 2018). While there is decent agreement between human and automatic judgments on bad model outputs, disagreements tend to be substantial on good outputs. Our analysis provides another example in support of these observations.

8 Conclusion

Question Answering on narrative text is a major challenge for the current methods in NLP. While the NarrativeQA data set provides an excellent benchmark, it is comparatively small, and does not allow training of *extractive* question answering, an arguably more straightforward task compared to extractive Q&A. We heuristically constructed an extractive summary-level Q&A data set and showed that it can be used to train accurate sentence- and span-level answer extraction systems from summary text. We also applied our models to full book text and showed that it outperforms IR-based retrieval systems when incorporated in a entity classification network.

We believe that narrative QA necessitates availability of rich prior information and training signal. We incorporated prior knowledge through pre-trained BERT embeddings, and used heuristic but inexpensive data for supervised training. We believe that our approach opens up avenues for more sophisticated data creation methods, leveraging better the rich information in the full book text. We will follow this avenue in future work.

Acknowledgments

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sathish Reddy Indurthi, Seunghak Yu, Seohyun Back, and Heriberto Cuayahuitl. 2018. Cut to the chase: A context zoom-in network for reading comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 570–575, Brussels, Belgium. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918, Berlin, Germany. Association for Computational Linguistics.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Densely connected attention propagation for reading comprehension. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4906–4917. Curran Associates, Inc.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Hugo Zaragoza, Nick Craswell, Michael J Taylor, Suchi Saria, and Stephen E Robertson. 2004. Microsoft cambridge at trec 13: Web and hard tracks. In *TREC*, volume 4, pages 1–1.