Review of a Doctoral Thesis at FIT BUT

Jordan Boyd-Graber

February 28, 2024

It was my pleasure to review "AUTOMATED FACTOID QUESTION ANSWERING AND FACT-CHECKING IN NATURAL LANGUAGE", a thesis from Martin Fajcik. The work is technically sound, addresses real-world problems, and is well written. I strongly recommend its acceptance, as it is an excellent contribution to the scientific literature.

1 Doctoral Thesis

1.1 Appropriateness and Relevance

Question answering and fact checking are key tasks in natural language processing and the techniques in this thesis are relevant more broadly to the development of artificial intelligence techniques used across computer science.

1.2 A summary of the Contributions of the Thesis

Joint Start/Stop Probability The most important chapter of the thesis carefully examines the formulation of the probabilities of the start and stop spans for answers in extractive question answering. The chapter brilliantly shows that when viewed independently, systems can latch on to contradictory clues and extract overly long (or short) information that are not actually good answers. This helped me better understand extractive question answering and is an important contribution to future models.

My only suggestions would be:

1. I appreciated the qualitative discussion (Table 3.7 was very helpful) that the spans were shorter, and the histogram of the lengths. However, this would have fit together well with the length heuristic discussion if it were presented a little earlier.
2. I found the notation of 3.3 and 3.4 a little confusing. While the marginal probabilities use auxiliary objectives, it's unclear which parameters are shared and how. Is θ a single vector, does it only apply to the H representations, or something else?
3. I would have appreciated a more theoretical discussion of Pareto optimality. Usually, when used in theoretical settings, it means that the choice is robust under all settings, but the analysis here seems to be wholly emperical.

Redundancy This chapter shows that many of the answers to questions appear multiple times in source corpora. This is an example of some of the most impressive research: something that seems obvious in hindsight but not previously discovered.

That said, the contribution would have been stronger with more analysis of the pruner and what exactly it's learning. I could see a story that it's learning to find the most information-dense passages or that it's finding the most trustworthy pages. However, given the emphasis on NQ, I could also see that it's just recreating the preferences of the Google search engine. I would have appreciated the same level of qualitative analysis as for the answer span chapter.

Efficient QA I remember R2D2 from the Efficient QA competition in 2020. I was impressed then and remain impressed by its ability to answer most of the questions in the competition despite significantly compressing the underlying repository of documents. That it did so well despite not having the resources of the other industrial teams. I think with the Docker hacking and model compression that Facebook was able to do, R2D2 could have been even more competitive in the results. I appreciate the comparative analysis of the various submissions (which was presented better than in in our official Efficient QA report).

Rumour Stances While the overall thesis is strong, I felt that this was the weakest chapter. The story and motivation leans too much on the shared

task, and while the overall design is reasonable, and I see not obvious avenues for improvement, I'm not sure that I learned anything from this chapter either about text modeling or about the problem of rumor stance detection. Perhaps adding more qualitative examples would help.

Fact Checking This chapter tackles an important problem—fact checking—by taking existing models, questioning the underlying assumptions, and then building up a new model that not only selects evidence but also provides signals to users of which evidence is relevant and important.

However, I think the presentation of the model could be more straightforward: the motivation and the modeling depends on the latent variable formulation, so I think a more compelling presentation would start by first motivating the latent variable approach, giving an intuition of what the latent variable means, and then going into the math. Back in the old days of statistical NLP, a rule was that you typically wanted to have English names for any latent variables you introduce both the ease references and to improve intuitions.

Given the emphasis on interpretability at the start of the section, I also would have liked an evaluation that could explicitly answer whether the results are useful. E.g., something along the lines of the experiment of Fool Me Twice, where users need to explicitly decide if a statement is true or not when the evidence is presented incrementally. If I'm understanding the presentation of the evidence, everything is presented all at once, so there's no guarantee that you've presented the minimal necessary information to validate a claim. For example, there's some recent evidence that retrieved information is less efficient than generated explanations from an LLM,[1] and users might be unreliable in rating which is more important.

1.3 Novelty and Significance:

Again, this is a very strong contribution to the literature, with potential ability to help shape the training of larger language models (e.g., given the redundancy for question answering, perhaps similar approaches could be used to train large language models with less data) and in creating useful mechanisms for Internet users to better spot disinformation online.

[1] `https://arxiv.org/abs/2310.12558`

1.4 Evaluation of the Formal Aspects of the Thesis:

This thesis is technically sound and well written. There are only minor issues, which I list below by page number (missing word denoted with *italics*, grammar error with underline, text to remove in [brackets])

10 Wang and Jiang should be inline citation

11 Most style guides prefer footnotes after punctuation

12 For the notation "given", most texts use $p(a \mid b)$, not $p(a|b)$ (but better to create a macro)

```
p(a\,\mid\,b)
```

13 Hence the model produces *an* obviously wrong answer.

13 I've usually seen it written as reader-retriever (which now that I write it doesn't make sense since the retriever comes first)

13 Some work (Nogueira and Cho, 2019; Luan et al., 2021) further adopts *a* computationally expensive reranking step on (or remove "step")

14 Moreover, other work uses *an* abstractive reader

14 Here, the thesis contributes in: ("makes contributions in" or "contributes")

14 How much of this set can we prune out, without damaging the system's performance?. (Remove final period)

15 *the* number of relevant evidences can be significantly larger than k

20 context-token (should be written with en–dash)

21 "translate English to German: That is good." and decoder input "Das ist gut.". (Issues with punctuation)

22 Gradient Accumulation is a memory-computation trade-off technique, which allows training model with a larger minibatch size[,] than the number of samples that fit into memory.

22 iteration[,] and summing

22 GPU[']s G

24 requires beam search, which slows down

24 This may be caused by the belief[,] that enumerating

26 Inline equation after 3.2 goes into column

26 of *the* span's start

26 for *our* language representation model

30 Use mbox for f_{MLP} so the kerning is correct

38 Years 2020-2021: Should be "The period between 2020–2021" (note the en dash)

39 composed of *a* DPR passage retriever

39 from *a* autoregressively factorized probability space

39–41 The ground truth passage—annotated the same way as in Karpukhin et al. (2020)—is primarily used as a positive sample (add em dash)

41 *The* exractive reader estimates

41 Equations 4.4–4.7 should be align environment justified around equal sign (e.g., &= in each line)

69 Before submission, we trained 100 models differing *only in* their learning rates

1.5 Quality of Publications:

Much of the work presented in the thesis has been published in peer-reviewed workshops, but two keystones (R2-D2: A Modular Baseline for Open-Domain Question Answering and Claim-Dissector: An Interpretable Fact-Checking System with Joint Re-ranking and Veracity Prediction) appeared at ACL Anthology Findings, a top-tier venue. Moreover, R2-D2 also competed in the Efficient QA competition, which was a non-traditional publication venue. I assume that others are pending review.

2 Overall R&D Activities Evaluation:

The thesis represents clear evidence of the ability to formulate a line of research and execute it. Moreover, it represents a triumph over challenging external circumstances: the Corona pandemic at the start of the theis, the sea change of GPT upending existing research, and the explosion of new methods and resources it unleashed.

Were there a good project fit, I would consider the candidate for a post-doctoral at my university.

3 Conclusion

In my opinion the doctoral thesis and the student's achievements meet the generally accepted requirements for the award of an academic degree.

College Park, Maryland; United States 28. Feb 2024

Jordan Boyd-Graber