Beyond the Hype: Making Progress on Natural Language Systems

Marc van Zee

I feel honored to contribute to this volume celebrating Leon's 50th birthday. From 2013 to 2017 I've been one of Leon's PhD students at the University of Luxembourg, working on logics of intention and formal argumentation. Though my academic career is still in its infancy, there are a few people I met who I consider to be special, and Leon is one of them. Perhaps it is because we are both Dutch, which makes me enjoy his direct approach. We've had many passionate discussions, more often than not in a bar while have a beer. I am sure a neutral observer would think two friends were having a quarrel, but the truth was far from it. I enjoyed those discussions a lot, and I'm quite sure Leon did as well. Professionally I've learned a lot from him, and he played an important role in shaping many (though not all) of my attitudes towards research. What I admire most in Leon is his ability to combine broad intuition with a mastery of technical details. Since this piece is too short for a lot of technical details, I've chosen to present some broad intuitions instead. The main claim is that a lot of progress can still be made with natural language systems as commonsense reasoners, despite sensational claims in the media. I propose we develop a benchmark dataset for commonsense reasoning, which would allow us to better evaluate natural langauge systems.

Sensational Media Reports on Natural Language Systems

The trend of sensational 'click-bait' headlines, excessive media attention to minor events, and 'fake news' seems to trickle down to all parts of our society. Politics, sports, the business world, and recently AI research as well. Current media report on AI more and more drift away from its purpose, which is to inform the public in a neutral and understandable way. The extent to which scientific results are picked up, bended and exaggerated is harmful because it creates a disconnect between what the common person expects from AI, and what it actually can do. Let me give you two examples.

Chatbots inventing their own language In August 2017, various news sources (such as the Mirror[1], the Sun[2], and the Digital Journal[3]) published articles with headlines similar to "Researchers shut down AI that invented its own language". The articles discussed an experiment conducted by Facebook researchers who were teaching chatbots how to negotiate with each other. As reported, the researchers discovered during tests that the bots managed to create their own machine language spontaneously. According to the article in the Sun: "UK Robotics Professor Kevin Warwick said: 'This is an incredibly important milestone, but anyone who thinks this is not dangerous has got their head in the sand. 'This is the first recorded communication but there will have been many more unrecorded. 'Smart devices right now have the ability to communicate and although we think we can monitor them, we have no way of knowing.'"

The reality is somewhat less dramatic. The two bots mentioned above were designed, as explained in a Facebook Artificial Intelligence Research unit blog post in June[4], for the purpose of showing it is "possible for dialog agents with differing goals (implemented as end-to-end-trained neural networks) to engage in start-to-finish negotiations with other bots or people while arriving at common decisions or outcomes."

The only thing these bots were doing was deciding together how to split a list of given items (such as books, hats, and balls). Facebook did indeed shut down the conversation, but not because they feared they had created a potential HAL 9000. FAIR researcher Mike Lewis told FastCo they simply decided "our interest was having bots who could talk to people," but not to each other. In the current game for content and attention, this story evolved from a measured look at the potential short-term implications of machine learning technology to sensational doomsaying.

AI outperforms humans in natural language understanding In January 2018, the Newsweek published an article titled "Robots can now read better than humans, putting millions of jobs at risk".[5] Researchers from Microsoft and Alibaba separately claimed that their AI software is a good as, if not better than, humans at understanding the written word.[6] Luo Si, chief scientist for natural language at Alibaba's institute of Data Science and Technologies stated: "It is our great honor to witness the milestone where machines surpass humans in reading comprehension." The core of this achievement is the Stanford Question Answering Dataset (SQuAD)[7] containing over 100,000 pieces of text from more than 500 Wikipedia articles. The dataset contains pieces of text from Wikipedia followed by a set of five questions and answers. For example, from a page

[1] https://www.mirror.co.uk/tech/robot-intelligence-dangerous-experts-warning-10908711

[2] https://www.thesun.co.uk/tech/4141624/facebook-robots-speak-in-their-own-language/

[3] http://www.digitaljournal.com/tech-and-science/technology/a-step-closer-to-skynet-ai-invents-a-language-humans-can-t-read/article/498142

[4] https://www.facebook.com/dhruv.batra.dbatra/posts/1943791229195215

[5] http://www.newsweek.com/robots-can-now-read-better-humans-putting-millions-jobs-risk-781393

[6] https://blogs.microsoft.com/ai/microsoft-creates-ai-can-read-document-answer-questions-well-person/

[7] https://rajpurkar.github.io/SQuAD-explorer/

on southern California[8], a text chunk has the question "what is the name of the border to the south?". The correct answer is "the Mexico-United States border."

This all looks very impressive, until one reads the description of SQuAD more closely. It consists of "questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage."[9] In the end, the only thing the AI has to do is to figure out that it needs to locate the words relevant to "border" and "south," and extract "Mexico-United States border." from the original text. Although this is an impressive achievement, can we really call this a breakthrough in machine intelligence? I argue not. In fact, the computers don't seem to have understanding of the *meaning* of the words. Roughly stated, the software transforms the text to similarity matrices, making it easy for algorithms to find the answer in the text by searching for related or matching words using these vectors.

Some Commonsense Reasoning Questions

One of the consequences of media report such as those above is that it is very difficult for the average person, and even for a researcher in the field, to determine how well computers actually are performing at understanding text. Moreover, many of the dataset, such as the SQuAD dataset I mentioned above, are specifically focussed on machine-learned approaches, and therefore often focus on a sub-problem in which machine learning can book progress.

Query	**S**	**A**	**G**
My brother is Jeroen. Who is my brother?	✓	✗	✓
My brother likes football. What does my brother like?	✗	✗	✗
My brother lives in Amsterdam. Where does my brother live?	✓	✗	✗
My father is married to Wilma. Are my parents married?	✗	✗	✗
I want to go skiing in Ischgl, Austria coming Friday with three friends. How should I get there?	✗	✗	✗
I want to be in Zurich at 10pm tonight. What time should I leave?	✗	✗	✓
I ate a sandwich and an apple. How many things did I eat?	✗	✗	✗
Remember that two friend are coming for diner tonight. How many people are coming for diner tonight?	✗	✗	✗
I ate rice last night. What did I eat?	✗	✗	✗

Table 1: Some commonsense reasoning queries for three state-of-the-art natural language systems: Apple's Siri (**S**), Amazon's Alexia (**A**), and Google's Assistant (**G**). A checkmark (✓) means the answer is correct, a crossmark (✗) means the answer either is incorrect, or not understood.

To get an idea of how good current natural language systems are at commonsense

[8] `https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Southern_California.html`

[9] This is noted by a blog post by Katyanna Quach at The Register: `https://www.theregister.co.uk/2018/01/17/ai_human_reading/`

reasoning, I've constructed a set of 9 very simple queries, each query consisting of one or more sentences (Table 1, column 1). Note these queries were not carefully hand-picked by me, I simply constructed them randomly by making use of some well-known commonsense reasoning domains, such as time and action, planning, space, interaction between people, and some natural language constructs such as quantification and numbers. These queries should be easy to understand for any child older than 12 years old, and require no expert knowledge. However, they do require some understanding of the text, meaning the answer can (often) not simply looked up in the previous sentence. I've tested these queries on three state-of-the-art natural language systems: Siri by Apple (**S**), Alexia by Amazon (**A**), and the Google Assistant (**G**).

The results show that all systems have a very limited understanding of most of my queries. It seems there is some support for family relations and naming, for instance, Siri is able to recognize "Jeroen" as my brother in the following conversation:

Q: My brother lives in Amsterdam
A: I currently don't have Jeroen's address. Click here to edit it
Q: Where does my brother live?
A: Here is the home address for Jeroen

Sometimes a system seems to understand the sentence, but it is often simply copying previous parts of the conversation (similar to the SQuAD example above):

Q: My father is married to Wilma
A: Ok, I'll remember that
Q: Are my parents married?
A: I remember you told me: "My father is married to Wilma"

Unfortunately, in most other cases the systems either do not understand the query, or return results from a search machine.

Towards a benchmark dataset for commonsense reasoning

It is important to stress that the question set in Table 1 is completely arbitrary and can by no means be used as a scientific result. The questions are not categorized in any way and the number of queries is far too low to be of any value. Moreover, it seems that most state-of-the-art natural language systems are not equipped to answer the kind of questions I have posed here, so it may come of no surprise that they fail to answer almost any question correctly.

The only point I want to make here is that there seems to be a disconnect between the media reports on natural language systems, and what they actually can do. Sensational headlines make it appear that we are at an inflection point in history, but a simple experiment gives a different view. The problem is that we currently don't seem to have a dataset that could give a voice to this lack of understanding in natural language systems. We have tons of dataset specialized to machine learned approaches, and extrapolating from progress in those datasets makes it appear as if we are making progress on natural language in general.

My suggestion is to create a benchmark dataset of commonsense reasoning. A useful starting point may be Ernest Davis' recent survey on logic-based commonsense

reasoning approaches [1]. He outlines various domains of commonsense reasoning, including time, space, physical, knowledge and belief, folk psychology, and interaction between people. Not only does such a dataset seem interesting from a scientific point of view, it may be a necessary wake-up call for AI research, and more importantly to the public, that we are by no means close to developing true intelligence yet.

References

[1] Ernest Davis. Logical formulizations of commonsense reasoning: A survey. *Journal of Artificial Intelligence Research*, 59(1):651–723, 2017.