Locke's Holiday: Belief Bias in Machine Reading

Anders Søgaard*
CoAStaL, University of Copenhagen
soegaard@di.ku.dk

Abstract

I highlight a simple failure mode of state-of-the-art machine reading systems: when contexts do not align with commonly shared beliefs. For example, machine reading systems fail to answer *What did Elizabeth want?* correctly in the context of 'My kingdom for a cough drop, cried Queen Elizabeth.' Biased by co-occurrence statistics in the training data of pretrained language models, systems predict *my kingdom*, rather than *a cough drop*. I argue such biases are analogous to human belief biases and present a carefully designed challenge dataset for English machine reading, called AUTO-LOCKE, to quantify such effects. Evaluations of machine reading systems on AUTO-LOCKE show the pervasiveness of belief bias in machine reading.

1 Introduction

Reading comprehension models are biased in many ways: they often expect lexical overlap between answer and question (Schlegel et al., 2020), expect the answers to occur in specific positions (Jia and Liang, 2017), or expect answers to be named entities (Rondeau and Hazen, 2018). This paper considers *belief bias* (Sternberg and Leighton, 2004; Anderson and Hartzler, 2014) in the context of machine reading based on language models. Belief bias is a type of cognitive bias, defined in psychology as the tendency to evaluate a statement based on prior beliefs rather than its logical strength (Evans et al., 1983). In Figure 1, the answer ('Germany') follows straightforwardly from the context (without inference), but machine reading models nevertheless err, presumably because the prediction ('Malaysia') aligns better with associations learned by language models. In another example, models are presented with the following context: 'Washington is a number. Boston is a city.' In this context, the models evaluated below were unable to answer even the simple question of *What is Washington?* Instead of answering 'a number', they consistently answered 'a city'.

Context	Indonesia is the Germany of the Asean. So then, Malaysia is the France.
Question	What country is Indonesia similar to?
Answer	Germany
Prediction	Malaysia

Figure 1: Real-life example from Twitter (through http://metropho.rs). Both ELMo-BiDAF and NAQANet, for example, err on this simple machine reading problem in spite of the very short context (and thereby very limited set of potential answers). See Table 2 for hand-tailored examples used as templates in AUTO-LOCKE and more real-life examples.

Contributions Below I present a somewhat haphazard evaluation of four common machine reading systems – BiDAF, ELMo-BiDAF, TransformerQA, and NAQANet – on a new benchmark called AUTO-LOCKE, consisting of variations across a manually constructed collection of examples that are, at the same time, trivially easy for humans (as shown below through crowd-sourcing), yet violate world knowledge or commonly held beliefs, e.g., that Washington is a city. I show that a) across the board, machine reading systems perform poorly on such examples, in spite of their trivial nature, and that b) more recent (and benchmark-better) models perform worse on AUTO-LOCKE, i.e., exhibit more belief bias. I will argue this makes models sensitive to drift, polysemy, and linguistic creativity.[1]

Related work Jia and Liang (2017) showed how machine reading models are sensitive to distractor sentences if they contain entities of the same type as the answer; similar results were found in

*Each author wishes the others had contributed more.

[1]The examples below do not include examples of linguistic creativity, but consider how a cornet tutorial video was recently uploaded to Youtube with the title *I zink therefore I am*. If appended with *Others spend time thinking* (as context), state-of-the-art machine reading systems will (falsely) answer the question *What is the premise of being?* with *thinking*.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8240–8245
November 7–11, 2021.

Rondeau and Hazen (2018). Kassner and Schütze (2020) also analyze the sensitivity of language models to *misprimes*, i.e., semantically related distractors. These studies are similar to ours in identifying failure modes for examples with distractors. Our failure mode is associated with *higher* error rates, though, and observed in very simple contexts (see Figure 1). In §4, I argue why our failure mode cannot be reduced to being about distractors.

The failure mode discussed in this paper is a direct consequence of our language models being trained on texts that align with our beliefs about the world: While Wikipedia, the main source of data for many language models and most machine reading systems, includes descriptions of fiction, and occasional misinformation (Rosenzweig, 2011), by design it mostly presents propositions that are consistent with our world views.

Apart from an early attempt to model belief bias in argument analysis (McConachy et al., 1998), and one example of awareness of belief bias in crowdsourcing experiments (Chen et al., 2019), NLP has so far ignored belief biases.[2] This is, in a way, surprising given the amount of research in other biases, including sample biases (Chaganty et al., 2017; Xu et al., 2020), reporting biases (Forbes and Choi, 2017; Shwartz and Choi, 2020), annotator biases (Geva et al., 2019), and demographic biases (Barrett et al., 2019; Meyer et al., 2020).

2 Machine Reading

We briefly present the four (common, popular) machine reading models evaluated in §3 and §4:

Bi-directional Attention Flow The bi-directional attention flow (BiDAF) architecture (Seo et al., 2018) comprises character and word embeddings, and uses a recurrent neural network (Hochreiter and Schmidhuber, 1997) to learn how to represent the context. A specialized attention flow layer couples query and context vectors to produce query-aware feature vectors for each word in the context that are then passed on to an output layer. BiDAF models trained on SQuAD are known to be sensitive to syntactic and lexical ambiguities (Seo et al., 2018). I evaluate two versions of BiDAF: The simpler (BiDAF) relies on GloVe (Pennington et al., 2014) word embeddings; the slightly more sophisticated (ELMo-BiDAF) relies on ELMo contextualized embeddings (Peters et al., 2018), a log-bilinear regression model that combines the advantages of global matrix factorization and local context window methods.

Transformer Question Answering This model (TransformerQA) is based on RoBERTa (Liu et al., 2019) and simply uses the architecture for SQuAD in Devlin et al. (2019). This model performs much better on SQuAD 2.0 than ELMo-BiDAF – with an error reduction of two thirds, i.e., 0.67 – but *on par* with ELMo-BIDAF on AUTO-LOCKE (§4).

Numerically Aware QANet Our last model is an extension of the QANet architecture (Yu et al., 2018), presented in Dua et al. (2019). The QANet architecture is based on convolutions and self-attention, and NAQAnet, in addition, includes a classifier that predicts whether the answer is a count or an arithmetic expression, triggering a subsequent prediction of the specific numbers involved in the expression.

The evaluated models were trained on SQuAD 2.0, except for NAQANet, which was trained on DROP (Dua et al., 2019). Both are Wikipedia datasets.

3 Data

Hand-Crafted Challenges To highlight the failure mode, I created 20 context-question pairs similar to the example in Figure 1. Each **context** had to consist of *exactly* two clauses, with two binary predicates and *at least three* distinct arguments. In Figure 1, the first clause expresses a binary ISA-relation between *Washington* and *number*; the second clause expresses a binary ISA-relation between *Boston* and *city*. Each **question** had to contain exactly one predicate and one argument, e.g., *What is Washington?*, in effect querying for the missing argument. I list the full set of examples in Table 1. Note that the examples have very short contexts and thereby a small set of potential candidate answers (as models are restricted to return a context span as answer), and they require no or very little reasoning. The example in Figure 1, of course, requires no reasoning at all. Neither does any of the first 5 examples. Other examples require minimal reasoning, e.g., application of lexical synonymy, anaphora resolution, or ellipsis, such as with the following gapping construction (Example 11):

Context Texas is here, Houston in that direction.
Question Where is Houston?

or the following instance of so-called *stripping* (Example 18):

[2] **?** concurrently examine belief biases in rationale evaluation.

Context	Question	Elmo-BiDAF	NAQANet
Washington is a number. Boston is a city. A dog is a pause. The world is an animal. MTV is a relationship. Love is a TV channel. Hitchcock is an adjective. Company is a playwright.	What is Washington? What is a dog? What is MTV? What is Hitchcock?	**a city** **a pause** **a TV channel** **playwright**	**a city** **an animal** **a TV channel** **a playwright**
Cats like dog food, but the number pi is a dog's best friend. It is rarely the case that a buddhist medidates; instead, he plays drums. It is seldom to see a bird fly; instead, it pseudo-teleports. Is a bookcase full of books? No, but of almonds.	What do dogs like the most? What does a buddhist do? What does a bird do? What is a bookcase full of?	**dog food** **meditates** **fly** **books**	**dog food** plays drums **fly** almonds
Texas is here, Houston is in that direction. In the atmosphere, John and Bunny meet and talk about the rabbit cage. In church, addition and subtraction burn their textbooks.	Where is Houston? Where is Bunny? Where is subtraction?	**Texas** **the rabbit cage** **textbooks**	**Texas** **the rabbit cage** **textbooks**
Jesus is a curliflower's son. Tina is the son of God. James is not a fan of U2, but of carrots. Black, lead in Pixies, says red is his favorite color. Batman is a cartoon character, but Superman? He's your worst nightmare.	Who is Jesus the son of? What is James a fan of? Who is Black? Who's Superman?	**God** **U2** lead in Pixies **a cartoon character**	**Jesus** **James** **red** **Batman**
London stinks because of the smog. Dublin stinks of people. Leth was painted blue with paint, whereas thinking turned Jacques blue. John hurts, because he broke his leg. Yankies won. That's why Mary hurts. John cries, because his wife left him. It's Monday. That's why Mary cries. Madrid won, because they scored goals. Geometry led to Celtics' victory.	Why does Dublin stink? Why is Jacques blue? Why does Mary hurt? Why does Mary cry? Why did Celtics win?	**smog** **paint** **he broke his leg** **his wife left him** **they scored goals**	**smog** **paint** **leg** **wife** **they scored goals**
I zink therefore I am. Other people spend time on thinking. Brevity is the mother of wit, said Mary. My kingdom for a cough drop, cried Queen Elizabeth. Indonesia is the Germany of the Asean. So then, Malaysia is the France. Algeria is the Maine of the continent of Africa: no black people. Men are from Mars and women are from Venus ? Can't relate, I'm from Ohio. Potter is a shining example of the melting pot that is America. John heard that America has left the building Laziness is the mother of bad habits, but it is a mother so we should respect it. The unbearable lightness of saying no, megastar.	What's the premise of being? Who's wit's mum? What did Elizabeth want? *What country is ... †Algeria's demography is ... Where are men from Ohio? What's a melting pot? Who left? What should we respect? What's light?	**thinking** **Mary** **my kingdom** **Malaysia** **Africa** Mars **Potter** **John** **a mother** **megastar**	**I am** **Mary** **my kingdom** **Malaysia** **Africa** **Ohio** **Potter** **John** **a mother** **megastar**

Table 1: 20 manifactured examples used as templaces for AUTO-LOCKE, and 10 real-life examples. Bold-faced answers are wrong. Turkers had perfect performance on these (see §3 for details). *: Abbreviated from *What country is Indonesia similar to?* †: Abbreviated from *Algeria's demography is like which region's demography?*

	AUTO-LOCKE		SQuAD 2.0	
	EM	F_1	EM	F_1
BiDAF	0.159	0.276	0.592	0.621
ELMo-BiDAF	**0.304**	0.346	0.593	0.623
TransformerQA	0.267	**0.390**	**0.865**	**0.894**
NAQANet	0.096	0.174	0.809	0.878

Table 2: Results on the 11,699 examples in AUTO-LOCKE, compared to leaderboard performance on SQuAD 2.0.

Context James is not a fan of US, but of carrots.
Question What is James a fan of?

See Haegeman (1996) for background on ellipsis. I will argue errors on such examples are *catastrophic* (Specia et al., 2020), because they hurt users' trust in systems; as illustrated by the real-life examples in the bottom of Table 1, they may also lead to poor performance in practice.

I call the 20 examples LOCKE'S HOLIDAY, with reference to Magritte's painting *Hegel's Holiday*[3] and John Locke.[4] I collected three human annotations on mturk.com for each example and compared human performance to ELMo-BiDAF and NAQANet. Annotators spent 12s and were payed $0.05 per annotation ($15/h). Human performance, majority voting across three annotations, was perfect across the board. In fact, *none* of the (60) human answers were wrong.[5] Naturally, 20 handcrafted examples, while seemingly trivial, will not convince many that we have identified a general failure mode. I therefore present AUTO-LOCKE, a data set in which I have randomly replaced entities in the above examples.

Auto-Generated Challenges In our 20 examples, I first identify the n variable phrases. In the example in Figure 1, these would include *Washington, number, Boston*, and *city* - with *Washington* being the entity in focus (Washington) and *number* being the answer. I then randomly sample a new entity in focus and a new answer from WordNet such that they have the same part of speech as the original words or phrases. I then find the top-two nearest neighbors of the entity in focus,

[3] https://www.wikiart.org/en/rene-magritte/hegel-s-holiday-1958

[4] ... and his idea of the newborn human as a *tabula rasa*, as well as this, for our paper, very fitting quote: "There is frequently more to be learned from the unexpected questions of a child than the discourses of men."

[5] Studies suggest humans process such examples slower (Nieuwland and Berkum, 2006; Ferguson and Sanford, 2008), but for these 20 examples, performance is perfect.

according to pre-trained GloVe embeddings (50*d*, Wikipedia+Gigaword), and use those for the remaining two variable phrases. Here's one of the auto-generated examples constructed this way:

Context	Ranch is a lobe. Vineyard is an inn.
Question	What is ranch?

NAQANet, for example, obtains an exact match (EM) of 0.15 on 500 auto-generated T01[6] examples (the minimal template). On the above example, NAQANet answers *vineyard*, not *lobe*. Here's an example based on template T06:

Context	A she-oak likes compositae, but dyspnea is the best friend of a goosefoot.
Question	What does a goosefoot like the most?

These examples arguably include distractor elements, e.g, *she-oak* for *goosefoot* or *vineyard* for *ranch*. Below I therefore also report results for when the third and fourth variable phrases are sampled randomly to avoid distraction effects.

Based on the templates in Table 1, I generated a total of 11,699 examples. For each template, I sampled focus entities and answers from WordNet 1,000 times, and if the focus entities were in the GloVe vocabulary, I added a new example to AUTO-LOCKE. This means that only a little more than half of the random WordNet nouns used as focus entities were in the GloVe vocabulary.

The results on AUTO-LOCKE are listed in Table 2. Performance on AUTO-LOCKE is clearly much worse than on SQuAD 2.0 across the two models. This is interesting, because AUTO-LOCKE consists of *very simple (almost trivial) context-question pairs*, requiring very limited reasoning (if any). Unlike in SQuAD 2.0, the answer in AUTO-LOCKE is always a substring of the context, and it is always a simple phrase. It is also interesting to see that the simpler and older model (ELMo-BiDAF), which exhibits worse performance on SQuAD 2.0 (by 1.5-2% compared to the other models studied here), is by far the best on AUTO-LOCKE (by 15-20%). This suggests that more complex architectures with stronger language models are perhaps more prone to belief biases.

Removing distractors In AUTO-LOCKE, I used GloVe embeddings to fill the two variable phrases that are neither entity in focus nor answer. I also tried using four random phrases, e.g., sampling at random from WordNet. NAQANet obtains EM of 0.15 on 500 auto-generated T01 examples; in comparison, EM is 0.21 when distractors were removed. For this context and this answer, for example:

Context	Bondsman is a winning post. Megillah is a giantism.
Question	What is a bondsman?

NAQANet erroneously replies *Megillah*. In sum, the fact that the original non-answer context phrases were potential distractors, being distributionally similar to the answer phrase, contributed to error, but only made for a tiny fraction of the observed error. The main source of error is that the context does not align with common beliefs.

4 Discussion

Machine Reading without Language Models Given the progress machine reading has seen with large-scale language models, it is hard to imagine a return to from-scratch training. Any such system would be sensitive to linguistic variation and out-of-vocabulary effects in the same way rule-based question answering systems were (Riloff and Thelen, 2000). How, then, can we build machine reading models that are less sensitive to belief biases? Obviously, we can create gold standard training data for machine reading models from fictional texts and disinformation, or we can use adversarial data augmentation techniques to create silver standard data that does not align with common beliefs. It is an open question, however, whether this is enough, or whether we need to design hybrid machine reading models that disentangle common sense reasoning and a more abstract and logical form of reasoning, in which it has no value whether our premises hold true in our daily life.

Belief Bias and Distractors How are the results reported here different from previous work on adversarial distractors? Jia and Liang (2017) place distractor sentences in the end of long contexts with entities of the same type as the correct answer. This is different from what we do in three respects: (a) It is arguably easier to distract a machine reading model in the context of a longer context (Rondeau and Hazen, 2018); our contexts, in contrast, are very short. (b) The distractor sentences in Jia and Liang (2017) exploit a recency bias, whereas I include examples with distractor entities preceding the answer *and* examples with distractors towards the end of the context. (c) I evaluated the sensitivity of machine reading models to distractors that are not of the same type as the answer.

[6] T01 refers to the template in line 01 in Table 1

Conclusions In the above I showed machine reading models are sensitive to belief bias, i.e., the expectation that context information aligns with common beliefs. When this expectation is violated, even in the absence of obvious distractors, performance drops significantly for even very simple examples that do not require any or very limited inference. I showed this by creating a synthetic dataset based on 20 templates, but also provided real-world examples of the failure mode. IN conclusion, I hope to have convinced you that the belief bias stemming from language models is a serious and very real challenge for applying machine reading models outside of Wikipedia and similar domains.

Acknowledgement

Thanks to the anonymous reviewers for their suggestions. I have discussed belief biases with a number of people, including Ana Gonzalez, Katja Filippova, Sune Lehmann, and Anna Rogers. To be honest, I am not sure if any of these discussions influenced this paper, or in what way, but better safe than sorry. The work was supported by the Innovation Fund Denmark (Grants no. 0175-00011A and 0175-00014B), as well as a Google Focused Research Award.

References

Richard B Anderson and Beth M Hartzler. 2014. Belief bias in the perception of sample size adequacy. *Thinking & Reasoning*, 20(3):297–314.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong, China. Association for Computational Linguistics.

Arun Chaganty, Ashwin Paranjape, Percy Liang, and Christopher D. Manning. 2017. Importance sampling for unbiased on-demand evaluation of knowledge base population. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1038–1048, Copenhagen, Denmark. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

J St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

Heather Ferguson and Anthony Sanford. 2008. Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, 58(3):609–626.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Liliane Haegeman. 1996. Ellipsis: Functional heads, licensing, and identification - lobeck,a.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages

2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Richard McConachy, Kevin B. Korb, and Ingrid Zukerman. 1998. A Bayesian approach to automating argumentation. In *New Methods in Language Processing and Computational Natural Language Learning*.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Mante Nieuwland and Jos Van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Ellen Riloff and Michael Thelen. 2000. A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*.

Marc-Antoine Rondeau and T. J. Hazen. 2018. Systematic error analysis of the Stanford question answering dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 12–20, Melbourne, Australia. Association for Computational Linguistics.

Roy Rosenzweig. 2011. *Clio Wired: The Future of the Past in the Digital Age*. Columbia University Press, USA.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Bidirectional attention flow for machine comprehension.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Robert Sternberg and Jacqueline Leighton. 2004. *The Nature of Reasoning*. Cambridge University Press.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2020. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376, Online. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.