Learning to Explain: Answering Why-Questions via Rephrasing

Allen Nie[1] **Erin D. Bennett**[2] **Noah D. Goodman**[1,2]
[1]Department of Computer Science [2]Department of Psychology
Stanford University
anie@cs.stanford.edu {erindb,ngoodman}@stanford.edu

Abstract

Providing plausible responses to why questions is a challenging but critical goal for language based human-machine interaction. Explanations are challenging in that they require many different forms of abstract knowledge and reasoning. Previous work has either relied on human-curated structured knowledge bases or detailed domain representation to generate satisfactory explanations. They are also often limited to ranking pre-existing explanation choices. In our work, we contribute to the under-explored area of generating natural language explanations for general phenomena. We automatically collect large datasets of explanation-phenomenon pairs which allow us to train sequence-to-sequence models to generate natural language explanations. We compare different training strategies and evaluate their performance using both automatic scores and human ratings. We demonstrate that our strategy is sufficient to generate highly plausible explanations for general open-domain phenomena compared to other models trained on different datasets.

1 Introduction

Allowing machines to provide human acceptable explanations has long been a difficult task for natural language interaction (Carenini and Moore, 1993). In order to provide explanations, systems need to acquire sophisticated domain-knowledge (Winograd, 1971), conduct causal reasoning over complex set of events (Hesslow, 1988) and over narrative chains (Chambers and Jurafsky, 2008), and apply commonsense knowledge (Levesque et al., 2011).

Past work has demonstrated that by leveraging human-curated structured knowledge bases such as WordNet (Miller, 1995) or ConceptNet (Liu and Singh, 2004), a system can learn to rank or choose between multiple plausible explanations and reach high accuracy (Luo et al., 2016; Sasaki et al., 2017). Recent successes have also shown that structured knowledge is not needed if one can train a language model on a large quantity of text. Such model can rank explanations based on the probability that each explanation might appear in natural text (Trinh and Le, 2018).

Phenomenon The city councilmen refused the demonstrators a permit because _____ ?
Original The city councilmen feared violence.

L2E-Seq2Seq (greedy):
They were not allowed to march in the city.
L2E-Seq2Seq (beam):
They did not have a permit.
LM-1B: They were not allowed to use the Cape Town airport.
L2W: It was the only thing in the city that could be done.
Open-Subtitle: I don't know.

Figure 1: We show the original Winograd schema sentence, the original offered explanation, and generated responses from our models.

While ranking explanations is an important task, the nature of explanation is more general than this. For one phenomenon, there might be many acceptable, natural, and useful explanations. In our work, instead of simply ranking or choosing explanations generated by humans, we propose to advance this important domain by directly generating the explanation. We measure success based on whether the generated sequence is grammatically correct and is a fluent, natural, and plausible explanation. This task has two advantages. First, it allows us to explore whether such a task is computationally feasible given the current learning framework. Second, answering open-domain

Proceedings of the 1st Workshop on NLP for Conversational AI, pages 113–120
Florence, Italy, August 1, 2019.

why-questions with plausible answers can make chitchat dialogue system more engaging, especially in response to "why" questions (which previous systems typically answer with degenerate responses such as "I don't know").

We show that simply training a language model on previously existing datasets is not enough. However, by leveraging dependency parsing patterns, we are able to construct two new datasets that will allow modern neural networks to learn to generate general-domain explanations plausible to humans. These new datasets of naturally occurring self-explanations (statements with "because", unprompted by a question) provide excellent training signal for generating novel explanations for a given phenomenon. We conduct human experiments on the important features that contribute to plausible explanations, and we describe a simple procedure that can rephrase **Why**-questions into a statement so our model can also function as a single-round chitchat chatbot that can answer **Why**-questions.

2 Learning to Explain

We use the discourse extractor developed by Nie et al. (2017). This extractor first filters sentences that contain a particular discourse marker (in our case, the marker "because"). It then uses predefined, pattern-based rules on the dependency parse obtained from the Stanford CoreNLP dependency parser (Manning et al., 2014) to split the sentence into two semantically complete sentence clauses, which can be referred as S1 and S2. Dependency parsing allows us to isolate explanations and phenomena from exogenous modifying phrases. Using these patterns to parse sentences with "because" also allows us to deal with the free order of the explanation and phenomenon in English. We formulate the **L2E** task as: given the phenomenon S1, the model needs to learn to generate a plausible explanation S2.

In addition to retrieving the phenomenon-explanation pair, we additionally retrieve five sentences that immediately precede the phenomenon to provide context. We concatenate the context with S1 using a special separation token, resulting in the sequence `C1, C2, ..., C5 <SEP> S1`. We hypothesize that context will allow the model to generate more thematically relevant explanations. We refer to this setting as the **L2EC** task.

Algorithm 1 Q-to-S1

Input: question q, dependency parsed.
Remove "Why". Start at the ROOT of q:
$subj$ = NSUBJ or NSUBJPASS
aux = first dependent in [AUX, COP, AUXPASS]
$vp^{(\text{lemma})}$ = all remaining dependents
if aux in ["do", "does", "did"] **then**
  vp = apply tense/person of aux to $vp^{(\text{lemma})}$
else
  $vp = aux\ vp^{(\text{lemma})}$
end if
$s = subj\ vp$

At last, we describe a procedure in Algorithm 1 that uses dependency parsing to turn **Why**-questions into the statement format of S1. This allows us to generate explanations as responses to **Why**-questions.

3 Model

3.1 Language Modeling

Language modeling focuses on modeling the joint probability of a sequence $p(X = x_1, ..., x_n)$. Using chain rule, this can be decomposed as $p(X) = \prod_{t=1}^{n} p(x_t|x_{<t})$, the product of conditional probabilities. The model parameterized by θ optimizes to maximize the log of the likelihood function $\mathcal{L}(X;\theta) = \sum_{t=1}^{n} p_\theta(x_t|x_{<t})$. In a neural language model, proposed by Bengio et al. (2003), a recurrent neural network is trained by truncated backpropogation through time to learn to model (theoretically) an infinitely long sequence.

3.2 Sequence to Sequence Modeling

First introduced by Sutskever et al. (2014), sequence-to-sequence (Seq2Seq) modeling estimates a conditional probability distribution of sequence Y given sequence X. $p(Y|X)$, where $X = \{x_1, ..., x_n\}$, and $Y = \{y_1, ..., y_k\}$. The overall objective function is similar to a language model: to maximize the log-likelihood of the probability of the Y sequence given the X sequence: $\mathcal{L}(Y, X; \theta, \psi) = \sum_{t=1}^{k} p_{\theta,\psi}(y_t|y_{<t}, X)$, with parameters θ for the encoder and ψ for the decoder. In our work, we experiment with different architectures for the encoder and decoder.

4 Data

We provide data accessibility statements in Appendix A.1 for each dataset we use to train and

evaluate our models. Our constructed dataset and web demo code are publicly available[1].

Source Dataset	Task	Data	Length
NewsCrawl	L2E	2.07M	29.4
NewsCrawl	L2EC	2.57M	149.4
Winograd	L2E	61	18.0
COPA	L2E	250	14.2
News Commentary	L2E/L2EC	6301	28.6

Table 1: Top are training datasets and bottom are evaluation datasets for each task. We report the average length of sentences for each dataset (S1 and S2 combined). News Commentary with context has 156.3 words on average.

4.1 Training Data

NewsCrawl Dataset We build up our training dataset from two large news datasets: Gigaword Fifth Edition (Parker et al., 2011) and NewsCrawl (Bojar et al., 2018). These two datasets contain news stories from 2001-2017, and are non-overlapping. We built our dataset of News explanation pairs using the pipeline described in Section 2 and then split into training, validation, and test. More details are reported in Appendix A.2.

BookCorpus BookCorpus is a set of unpublished novels (*Romance*, *Fantasy*, *Science fiction*, and *Teen* genres) collected by Zhu et al. (2015). We use a publicly available pre-trained BookCorpus language model from Holtzman et al. (2018). We refer to this model as **L2W**.

Language Modeling One Billion This dataset (LM-1B) is currently the largest standard training dataset for language modeling, roughly the same size as BookCorpus. This dataset is a subset of the NewsCrawl dataset, from 2007-2011. We use a pre-trained language model on this corpus from Jozefowicz et al. (2016). We refer to this model as **LM-1B**.

4.2 Evaluation Data

News Commentary (NC) Dataset We collect pairs from a public dataset that contains predominantly commentary written about current news[2]. We use this dataset as the main evaluation of the news-based explanation because 1). It is a separate dataset without any overlap with NewsCrawl; 2). This dataset still belongs to the same news domain, so it provides an in-domain evaluation for **L2E**, **L2EC** and **LM-1B** models.

Winograd Schema Challenge Subset (WSC-G) We use 61 example sentences in the Winograd Schema Challenge that contain the words "because" or "so". Similar to Trinh and Le (2018), we substitute the ambiguous pronouns with the correct referent and ask the model to generate the correct explanation "the trophy is too big" to the phenomenon "The trophy doesn't fit in the suitcase".

Choice of Plausible Alternatives (COPA) Roemmele et al. (2011) proposed a task that contains questions such as "The women met for coffee. What was the CAUSE of this?", and the model is asked to choose between two pre-defined causes. In our setting, we directly ask the model to generate a cause. For language models, we append "because" to the end of each COPA sentence and ask the model to generate the rest.

5 Experiments

5.1 Language Model Training

We use the same language model described in Holtzman et al. (2018). We train 10 epochs for both L2E and L2EC. We use a one layer LSTM (Hochreiter and Schmidhuber, 1997) with 2048 hidden state dimensions and 256 word dimensions. We chose these hyperparameters by tuning on the validation set of each task. Our language model achieved 51.64 perplexity on the L2E test set, and 37.61 perlexity on the L2EC test set.

5.2 Seq2Seq Model Training

We experiment with two architectures: LSTM encoder-decoder and Transformer (Vaswani et al., 2017). We find that with the L2E task, the Transformer architecture performed better, and for the L2EC task, the LSTM encoder-decoder performed better. We suspect that Transformer is worse when the source sequence is too long. We tune each architecture's hyperparameters extensively and we pick the best architecture for each task to evaluate on the evaluation datasets.

5.3 Automatic Evaluation

We use automatic metrics to evaluate the 8 models' performance on the News Commentary dataset. Even though this is a non-overlapping held-out dataset to our news training data, it is still

[1] https://github.com/windweller/L2EWeb
[2] https://www.project-syndicate.org/about

Data	S1	Generated S2	Rank	Reference S2	Rank
NewsCrawl	That banned his most threatening challenger, Rally leader Alassane Ouattara, from running for president because ___ ?	He was born in Burkina Faso.	—	He is only half-Ivorian.	—
NewsCrawl	The victim was only saved because ___ ?	He was wearing a seatbelt.	—	The dog turned on the former lifeguard.	—
NewsCrawl	I voted for George W. Bush because ___ ?	I thought he was the best person for the job.	—	That's the name you heard a lot of talk about.	—
WSC-G	An hour later John left because ___?	He didn't feel safe.	0.0	John promised Bill to leave.	0.67
COPA	The woman gave the man her phone number because ___?	She was too busy to be bothered by the man.	0.17	She was attracted to him.	0.5
NC	Moreover, ordinary Russians are becoming allergic to liberal democracy because ___?	They see it as a threat to their own interests.	0.16	Liberal technocrats have consistently served as window dressing for an illiberal Kremlin regime.	0.19

Table 2: **Example pairs** from our highest performing models with the original sentence as a reference. Human ranking score lower is better. We provide examples of especially poor-rated generations in the Appendix.

Model	L2E		L2EC	
	Acc	Perp	Acc	Perp
LSTM	36.2	41.4	**36.0**	**41.3**
Transformer	**38.2**	**33.1**	27.8	96.7

Table 3: We report the best per-token accuracy and perplexity evaluated for each tuned architecture on the L2E/L2EC validation dataset.

within the same domain. We find that L2E/L2EC based models obtained higher scores across all automatic metrics in Table 4. Our results also demonstrate that context matters for explanation. The L2EC task models, trained on context, can generate higher quality explanations than context-free L2E task models.

5.4 Human Evaluation

Ranking Explanations We evaluate the models' relative performance on generating explanations through a survey with human evaluators. 75 participants were recruited using Amazon's Mechanical Turk (AMT). Each evaluator saw 10 prompts from a single dataset, and ranked 7 to 9 explanations: the original explanation extracted from the dataset and the explanations generated by different models. 30 participants saw prompts from our Winograd dataset, 30 participants saw prompts from News Commentary, and 15 participants saw prompts from COPA. We report the results of this evaluation in the Human Ranking subsection of Table 4.

Rating Explanations In a followup survey, 60 human evaluators on AMT rated explanations generated by the L2E-Seq2Seq model with beam search and the original (between participants). Ratings were from 0 (extremely bad) to 1 (extremely good) along various dimensions of explanation quality. Results of this study are shown in Table 5. Generated explanations overall were rated worse than human explanations, but tended to be more good than bad (≥ 0.5) on all measures.

6 Discussion

The nature of phenomenon-explanation mapping has always been one-to-many. People can offer drastically different explanations to the same phenomenon. We argue that requiring the machine to generate plausible explanations is more useful and therefore a better goal for models to achieve. Models trained on traditional chatbot corpora are unable to answer why questions because of data sparsity. We note that the generated results are not similar to the original explanations but are often acceptable by human assessment.

Features of Explanations In the human rating experiment, our model was overall rated higher than the original explanations only on the grammaticality measure. However, this measure seems least representative of the overall explanation

Model	BLEU		ROUGE		METEOR		Human Ranking		
	Greedy	Beam	Greedy	Beam	Greedy	Beam	COPA	WSC-G	NC
L2E-Seq2Seq	**0.55**	0.37	18.8	18.3	7.4	7.6	**0.412**	**0.409**	0.454
L2EC-Seq2Seq	0.40	**0.47**	**19.9**	**19.7**	**8.6**	**8.8**	—	—	0.433
L2E-LM	0.25	0.20	15.9	16.8	6.1	6.7	0.515	0.572	0.479
L2EC-LM	0.36	0.38	17.0	17.7	6.7	7.3	—	—	**0.432**
LM-1B†	0.18	—	16.9	—	7.1	—	0.526	0.484	0.454
L2W†	0.00	0.00	14.0	13.9	6.7	6.8	0.511	0.523	0.625
L2WC	0.13	0.14	12.8	12.7	5.7	5.7	—	—	0.546
OpenSubtitle†	0.04	0.0	13.0	13.4	1.9	3.7	0.827	0.823	0.811
Reference	100	100	100	100	100	100	**0.266**	**0.238**	**0.267**

Table 4: BLUE, ROGUE, METEOR are evaluated on News Commentary test data. Any model with **C** in the name is evaluated with full context. Models with † are pre-trained models from other work. Only L2E-Seq2Seq uses the Transformer architecture, the rest LSTM. In human ranking, we report the average rank across participants. Top ranking is 0 and lowest ranking is 1.

	Original	L2E-Seq2Seq
Goodness	**0.699** [0.67, 0.72]	0.500 [0.45, 0.55]
Relatedness	**0.723** [0.70, 0.74]	0.590 [0.55, 0.63]
Grammaticality	0.684 [0.66, 0.71]	**0.738** [0.70, 0.77]
Helpfulness	**0.696** [0.67, 0.72]	0.512 [0.47, 0.56]
Plausibility	**0.710** [0.69, 0.73]	0.543 [0.50, 0.59]

Table 5: Results of rating study with human evaluators, average rating and bootstrapped 95% CI.

quality: ratings for most features were highly correlated with each other (0.771-0.865), but not with grammaticality (0.196-0.323). This shows that, while we can achieve plausible explanations with our models, more research is required in order to reach human-level quality.

Explaining as Generating Even though formulating the task of providing explanation as a sequence generation task allows us to leverage the rapid advancements in the natural language generation community, we sidestep a vast amount of literature that aims to provide informatively *correct* explanations as well as grounding explanations theoretically to the causal understanding of the situation (Halpern and Pearl, 2005). We also suffer from the same drawbacks noticed in natural language generation papers such as brevity and generic responses, failure to leverage long context, and being data hungry (Holtzman et al., 2018).

Exploring Linguistic Structures The curated dataset of explanation-phenomenon pairs provides an opportunity to explore descriptive structures and features of explanations. In principle, one can use this dataset to formulate frequent and common syntactic and semantic patterns for natural-sounding explanations. This would aid our understanding of how why-questions can be addressed satisfactorily.

7 Conclusion

We present the task of generating plausible explanations as an important goal for neural sequence-to-sequence models. We curate a large dataset of phenomenon-explanation pairs so that these models can learn to provide plausible explanations as judged by humans, and formulate responses to open-domain **Why**-questions.

Acknowledgement

We thank Barry Haddow and Michael Hahn for their advice and assistance. The research is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), via the Air Force Research Laboratory (AFRL, Grant No. FA8650-18-C-7826). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, et al. 2018. Proceedings of the third conference on machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Giuseppe Carenini and Johanna D Moore. 1993. Generating explanations in context. In *IUI*, volume 93, pages 175–182.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 56(4):889–911.

Germund Hesslow. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 11–32.

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

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning*, volume 46, page 47.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *KR*, pages 421–431.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Allen Nie, Erin D Bennett, and Noah D Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *arXiv preprint arXiv:1710.04334*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition LDC2011T07. *Philadelphia: Linguistic Data Consortium*. Web download file.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Data	S1	Generated S2	Rank	Reference S2	Rank
NewsCrawl	Perhaps the student was so good because ___ ?	He was so good at it.	—	Thursday wasn't the first time he acted as a moderator.	—
NewsCrawl	Cheaper prices drove computers out of the walls of these big companies because ___ ?	They were cheaper than the ones that were being used.	—	You suddenly didn't need all that money anymore.	—
WSC-G	Sam and Amy are passionately in love, but Amy's parents are unhappy about it because ___?	They don't want to be the first female president.	0.87	Sam and Amy are fifteen.	0.37
COPA	The hamburger meat browned because ___?	That's what it is.	0.83	The cook grilled it.	0.0
NC	The desperately poor may accept handouts because ___?	They are the only ones who can afford it.	0.83	They feel they have to.	0.12

Table S1: **Bad example pairs** from our lowest performing models with the original sentence as a reference. Human ranking score lower is better. Full list of WSC-G and COPA generations can be found in https://github.com/windweller/L2EWeb/blob/master/WinogradS2Generation.ipynb.

A Supplementary Materials

A.1 Data Accessibility Statement

The majority of the data we use are publicly available. We provide specific instructions on how to obtain these data below:

Gigaword 5th Edition This dataset is provided through Linguistic Data Consortium (LDC): https://catalog.ldc.upenn.edu/LDC2011T07. Even though this dataset is only available through subscription, most university libraries should have existing subscriptions, and only 20% of our training data comes from this dataset.

News Crawl Dataset The shuffled version of this dataset is publicly available[3]. We requested the original un-shuffled dataset from Barry Haddow[4] so that we can extract context for L2EC task. We believe this dataset can be easily accessed by the public upon an email request.

BookCorpus This dataset is no longer publicly available. However, there are many neural language models pre-trained on this dataset that are publicly available. We used one that can be accessed from https://github.com/ari-holtzman/l2w.

News Commentary Dataset This is also publicly available through the WMT workshop[5] similar to the NewsCrawl dataset. This dataset is not shuffled.

Winograd Schema Challenge The original version of this dataset is publicly available https://cs.nyu.edu/davise/papers/WinogradSchemas/WS.html. We use a processed version from Trinh and Le (2018), which can be accessed through Google Cloud Storage: gs://commonsense-reasoning/reproduce/commonsense_test/wsc273.json.

Choice of Plausible Alternatives This dataset is available at http://people.ict.usc.edu/~gordon/copa.html.

A.2 Training Data Curation

In order to automatically curate a sizable amount of training data, we choose large corpora that are made of news articles, due to the well-formedness of sentences and there are many phenomenon-explanation pairs in news stories. We use Gigaword fifth edition (Parker et al., 2011) which contains news stories from seven news agencies over the span of 2001-2010. We extracted paragraphs and tokenized the sentences. We discard non-English characters. Another large dataset of new articles comes from WMT-18, the NewsCrawl dataset (Bojar et al., 2018). This dataset spans from 2007-2017 collected from the RSS (Rich Site Summary) feed of 18 news agencies. The

[3] http://www.statmt.org/wmt18/translation-task.html

[4] http://homepages.inf.ed.ac.uk/bhaddow/

[5] http://data.statmt.org/wmt18/translation-task/news-commentary-v13.en.gz

only overlapping agency between Gigaword and NewsCrawl is Los Angeles Times. In addition t the randomly shuffled dataset we obtained fro the WMT-18 website, we additionally contacte the organization for the unshuffled version of dat We refer to this dataset as the NewsCrawl-ordere This dataset is slightly larger than the current r leased version of NewsCrawl and contains a cou ple of months of early 2018 data. We shuffle an then split both datasets into train/valid/test in sta dard 0.9/0.05/0.05. We use the validation and te set on this task to pick the best performing mode

A.3 Language Model Details

We use adaptive gradient descent (AdaGrad) with learning rate 0.1 and weight decay of 1e-6.

A.4 Seq2Seq Model Details

We built and trained our Seq2Seq model using OpenNMT (Klein et al., 2017). For the L2E task, we used a 6-layer Transformer model, with hidden dimension 512, feedforward layer dimension 2048, and 8 attention heads. We train with dropout

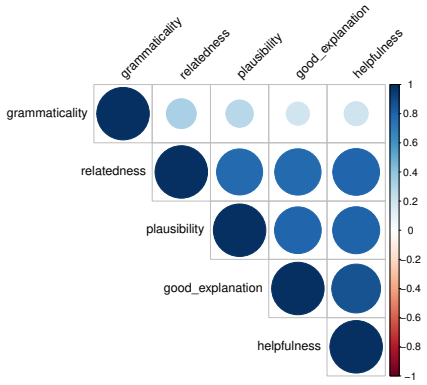

Figure S1: Correlations of human ratings on Winograd Schema Challenge explanations

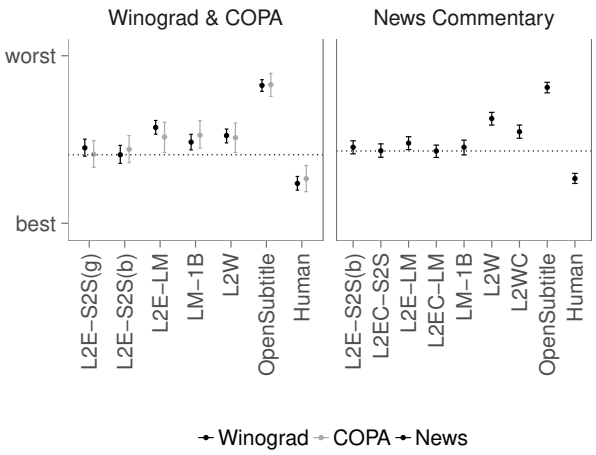

Figure S2: The average ranking of each model's generated response (lower is better).

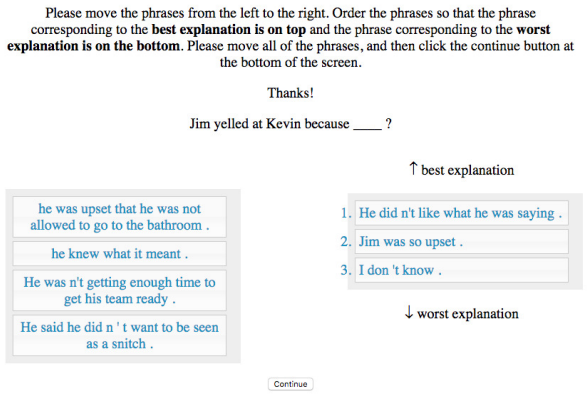

Figure S3: Screenshot of raking study.

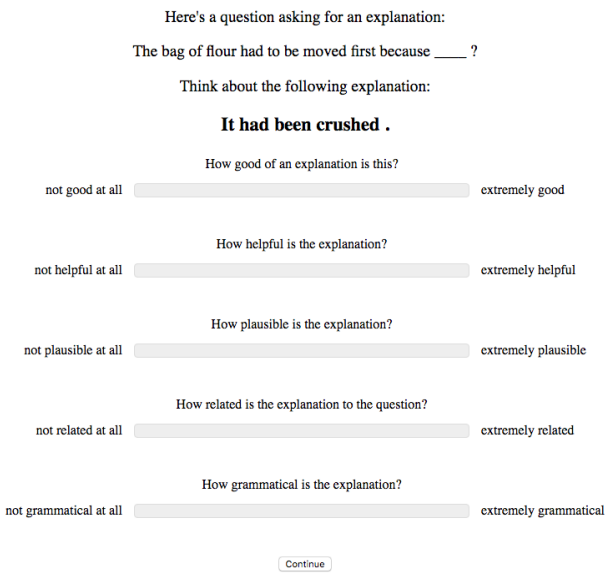

Figure S4: Screenshot of ratings study.