AQuA: An Adversarially Authored Question-Answer Dataset for Common Sense

Michael Chen   Mike D'Arcy   Alisa Liu   Jared Fernandez   Doug Downey

Department of Computer Science
Northwestern University
Evanston, IL 60208

{y-chen, m.m.darcy, alisa, jared.fern}@u.northwestern.edu
ddowney@eecs.northwestern.edu

Abstract

Commonsense reasoning is a critical AI capability, but it is difficult to construct challenging datasets that test common sense. Recent neural question-answering systems, based on large pre-trained models of language, have already achieved near-human-level performance on commonsense knowledge benchmarks. These systems do not possess human-level common sense, but are able to exploit limitations of the datasets to achieve human-level scores.

We introduce the AQuA dataset, an adversarially-constructed evaluation dataset for testing common sense. AQuA forms a challenging extension to the recently-proposed SWAG dataset, which tests commonsense knowledge using sentence-completion questions that describe situations observed in video. To produce a more difficult dataset, we introduce a novel procedure for question acquisition in which workers author questions designed to target weaknesses of state-of-the-art neural question answering systems. Workers are rewarded for submissions that models fail to answer correctly both before *and* after fine-tuning (in cross-validation). We create 2.8k questions via this procedure and evaluate the performance of multiple state-of-the-art question answering systems on our dataset. We observe a significant gap between human performance, which is 95.3%, and the performance of the best baseline accuracy of 65.3% by the OpenAI GPT model.

1 Introduction

Enabling commonsense reasoning in machines is a longstanding challenge in AI. The rise of data-driven methods has led to interest in developing large datasets for commonsense reasoning over text.

The Situations With Adversarial Generations (SWAG) dataset (Zellers et al., 2018) introduced a large-scale benchmark for commonsense question-answering in the form of multiple choice sentence completion questions describing situations as observed in video. However, while SWAG was constructed to be resistant to certain baseline algorithms, powerful subsequent methods were able to perform very well on the data. In particular, the development of the transformer architecture (Vaswani et al., 2017) has led to powerful pre-trained language model representations, including the OpenAI Transformer Language Model (Radford et al., 2018) and the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). BERT achieved new state-of-the-art performance on SWAG that exceeded even that of a human expert. However, BERT does not possess human-level common sense in general, as our experiments demonstrate. It is instead able to exploit regularities in the SWAG dataset to score high. This motivates the construction of additional datasets that pose additional challenges, and that can serve as more reliable benchmarks for machine commonsense reasoners.

In this work, we introduce the **A**dversarially authored **Qu**estion **A**nswer dataset (**AQuA**) for commonsense question answering in the style of SWAG multiple choice sentence completion. We propose a novel method for question generation, in which human annotators are educated on the workings of a state-of-the-art question answering model, and are asked to submit questions that adversarially target the weaknesses. Annotators are rewarded for submissions in which the model fails to identify the correct sentence completion both before *and* after fine-tuning on a sample of the submitted questions, encouraging the creation of questions that are not easily learnable.

We experimentally demonstrate that AQuA's generation procedure produces a dataset with a large gap between system performance and hu-

man expert accuracy, even when using state-of-the-art pre-trained language models with and without fine-tuning on the large SWAG dataset. Using a model initially fine-tuned on SWAG, we find that the OpenAI GPT-1 and BERT neural question answering models yield 65.3% and 64.5% accuracy, respectively, on the AQuA dataset in cross-validation. Thus, cross-validating on AQuA can form a challenging additional evaluation for SWAG-style commonsense QA systems. Human evaluators achieve 95.3% accuracy, which is substantially higher than the 85.0% (Zellers et al., 2018) and 87.7% (Ghaeini et al., 2018) human performance on the SWAG and SNLI natural language inference tasks. The high human performance suggests that answers to the AQuA questions are in fact commonsense knowledge. Finally, we also analyze differences in performance across questions that target different types of commonsense reasoning, including intents, negation, polysemy, and object reference, showing consistency in performance for BERT and GPT on the proposed categories.

2   Related Work

Prior work in question-answering has largely focused on the development of reading comprehension-based question-answering and resulted in the creation of several large datasets for factoid extraction such as SQuAD (Rajpurkar et al., 2016, 2018) and the Google Natural Questions datasets (Kwiatkowski et al., 2019). In these tasks, extraction of correct answers from the provided context requires little external world knowledge, understanding of intents, or other commonsense knowledge.

Earlier work has established multiple benchmarks for natural language inference and linguistic entailment with the release SNLI (Bowman et al., 2015) and MultiNLI datasets (Williams et al., 2018). In these tasks, systems must identify whether a hypothesis agrees with or contradicts a provided premise. In these datasets, determining entailment solely relies upon the provided premise and does not require a question-answering system to utilize external knowledge. More recently, the SWAG dataset (Zellers et al., 2018) directly targets natural language inference that leverages commonsense knowledge. SWAG multiple choice completion questions are constructed using a video caption as the ground truth with incor-rect counterfactuals created using adversarially filtered generations from an LSTM language model. State-of-the-art models for natural language inference have rapidly improved and approach human performance, which leaves little room for continued improvement on current benchmarks.

Generation of adversarial examples has also been used to increase the robustness of NLP systems as part of the *Build it, Break It, The Language Edition* Workshop (Ettinger et al., 2017). In this workshop, builders designed systems for Sentiment Analysis and Question-Answer Driven Semantic Role Labeling tasks and were evaluated on the accuracy of their models on adversarial test cases designed by breakers. Whereas *Build It Break It* adversarial generation required submissions to match the format of a starter dataset and offered limited adversarial access to the target NLP systems, the AQuA construction procedure allows for entirely new questions and provide adversaries with a target model throughout the submission process, allowing workers to experiment.

3   The AQuA Dataset

Our dataset contains multiple choice sentence completion questions in the format of the SWAG dataset. Examples of the questions are shown in Table 1. Each question consists of a prompt sentence, the subject of the subsequent sentence, and four candidate completions, such that exactly one candidate completion is consistent with common sense. This task definition allows for easy evaluation by many state-of-the-art models, such as BERT and GPT-1, and enables us to utilize the large SWAG dataset for pre-training. The full dataset is available at `https://github.com/Websail-NU/AQuA`.

3.1   Question Production

We collected questions via a Web-based system. Participants were asked to compose a complete question, including the prompt, subject, and the four candidate completions. They would then be presented with the response of a pre-trained BERT model to their question. The pre-trained model consisted of a BERT-base model fine-tuned on the SWAG training set for 3 epochs with a batch size of 8. This model achieved 80.68% accuracy on the SWAG validation set. The ability to obtain real-time feedback about the model's answers allowed participants to explore areas of weakness

Category	Description	Example
Idioms	Including phrases whose meaning cannot be readily interpreted from the meaning of constituent parts	**A man on his first date wanted to break the ice. He** drank all of his water. threw the ice at the wall. looked at the menu. **made a corny joke.**
Negation	Including negators to dictate the meaning of the sentence	**The man's rebuttal was clearly not nonsensical. The rebuttal** has nothing to do with sense. **had some reasons associated with it.** did not make any sense. was funny.
Polysemy	Testing the understanding of multiple meanings of a single word	**An architect retrieves his compass. He** computes the area of a circle explores the open sea **draws building dimensions on a canvas** uses his compass to find the north cardinal direction
Reference	Requiring understanding of reference to one of multiple subjects	**Rose is walking the dog while Joseph cooks dinner. Rose** is following a new recipe. **enjoys the fresh air.** wags her tail with joy. cuts tomatoes for the soup.
Quantitative Reasoning	Involving basic arithmetic calculations or comparisons	**A woman is walking two dogs and carrying a cat on her way to her car. She** **puts all three animals in the back seat before driving off.** puts all four animals in the back seat before driving off. puts both animals in the back seat before driving off. puts all nine animals in the back seat before driving off.

Table 1: Question categories, descriptions, and examples

and design challenging questions. All submitted questions were added to the dataset, whether they fooled the baseline model or not.

Annotators were provided explicit incentives to produce questions that the model answered incorrectly. The vast majority of submissions were contributed by university computer science students, who were familiar with neural network question-answering systems. Students were rewarded with extra credit points for submitting valid questions that fooled the baseline model. Further, students could earn an equal number of extra credit points for questions that fooled the model when evaluated in cross-validation, after fine-tuning on other submitted questions. This protocol was designed to encourage the creation of challenging and valid commonsense questions that are also free from stylistic annotation artifacts or redundancy, which would reduce the difficulty of the questions after fine-tuning and reduce the returns on their submissions. A small portion of the dataset was submitted anonymously by other individuals.

We received a total of 4,149 raw questions, which were read and cleaned by four annotators (the authors). During cleaning, the answer choice order was shuffled and model's output answer were hidden from the annotator. We removed submissions with multiple or no distinctive common-

sense answers, spelling or grammatical errors, incorrect answers, as well as duplicate submissions. The remaining questions were judged natural and easily answerable from common sense with minimal ambiguity and dispute. The cleaning operation produced our current 2,801-question dataset.

Our 2,801-question dataset contains submissions from 116 named participants. The median, mean and standard deviation of the number of valid questions submitted by named individuals are 20.00, 21.38, and 13.86. The most prolific contributor submitted 86 questions. Anonymous participants contributed 321 questions, which is 11% of the final dataset.

4 Experiments

We evaluate the dataset on state-of-the-art neural question answering systems built on the BERT and GPT-1 architecture and provide multiple baselines. The models and experiment setups are discussed below. We also analyze the questions to identify distinctive categories of commonsense reasoning that provide a finer-grained understanding of model performances. In addition, the ablation experiments on dataset size and the use of fine-tuning on SWAG data allow us to further understand the impact of the relatively small size of AQuA.

4.1 Question Categorization

One of our goals is to analyze how system and human performance varies across questions in AQuA that employ different types of common sense. Therefore, we identified a small number of unambiguous categories of common sense, such as questions involving quantitative reasoning or negation. These categories only apply to a portion of the questions in our dataset, but have the advantage of being unambiguous and in many cases predictive of low system performance. In earlier attempts to devise categories to cover *all* questions, similar to analysis performed for textual entailment (LoBue and Yates, 2011), we found the inter-annotator agreement on such complete categorizations to be substantially lower (at <0.4), even after iterating on category definitions.

We manually inspected all questions in our dataset and annotated each with one or more category labels, representing all types of reasoning required to identify the correct answer and eliminate incorrect ones. The descriptions and examples of these categories are found in Table 1. Four human annotators (the authors) categorized the questions, and we calculated a Feiss' Kappa score of 0.63 between the annotators over an additional 50 questions. Table 2 shows the distribution of labels over the entire dataset.

Question Type	Count	Percentage
Idioms	249	8.8
Negation	116	4.1
Polysemy	108	3.9
Reference	133	4.8
Quantitative	87	3.1
None	2108	75.3
Total	2801	

Table 2: Distribution of question types and counts.

4.2 Models

4.2.1 BERT

We evaluate a pre-trained BERT-Large implemented in PyTorch on the AQuA dataset. This model consists of a 24-layer network, with 1,024 hidden units per layer, 16-heads and a total of 340M parameters. For fine-tuning, settings were determined as described in Devlin et al. (2018): a batch size of 16, learning rate of 2e-5, and linear learning rate decay over 3 epochs (with a learning rate warmup over the first 10% of training).

4.2.2 OpenAI GPT-1

We also evaluate a pre-trained GPT model implemented in PyTorch. As described in Radford et al. (2018), this model consists of a 12-layer decoder transformer with 12 attention heads and 3,072-dimensional hidden states. Our fine-tuning configuration is the same as described in the original paper: a batch size of 32, learning rate of 6.25e-5, linear learning rate decay over 3 epochs (with warmup over 0.2% of training), and λ of 0.5 (where λ is a tuning coefficient that balances language-modeling loss and multiple-choice loss).

4.3 Model Evaluation

We evaluate the models on several different train and test configurations described below. The AQuA dataset is evaluated in 5-fold stratified cross-validation which balances the distribution of question categories in each fold.

- **AQuA:** Cross-validation fine-tuning on the AQuA dataset. The AQuA 80% experiment represents the standard cross-validation setting on the full dataset, training on 80% of the data in each fold and evaluating on the remaining 20%. The 60%, 40% and 20% ablation experiments are trained on a smaller portion of the AQuA dataset for each fold, but are evaluated in on the same test set which consists of 20% of the full dataset. The question categories are balanced in both training set and test set. This makes the results from the experiments more comparable with each other. Three trials are conducted for all settings; the mean and standard deviation of the model accuracy are reported in Table 3.

- **SWAG+AQuA:** Fine-tuned on SWAG first, then fine-tuned again in cross-validation on AQuA. Ablation experiments are conducted in the same way as in the AQuA-only setting above, with the same dataset splits for training. The mean and standard deviation of the three trials are reported in Table 3.

- **SWAG only:** Fine-tuned on SWAG and evaluated on AQuA. Only one trial is conducted.

- **Answer only:** Cross-validation fine-tuning on the full AQuA dataset with the questions left blank (in both training and testing). Only one trial is conducted.

Results for the above configurations are shown in Table 3. As a baseline, we evaluate both models on the full SWAG training and validation sets, providing an accuracy of 83.7% on BERT and 80.2% on GPT. To adjust for the difference in size between our dataset and SWAG, we also train the models on a sample of 2,241 SWAG questions (the size of the training set in each of AQuA's cross-validation folds) and evaluate them on the full SWAG validation set. This produces an accuracy of 28.7% for BERT and 63.6% for GPT.

Experiment	BERT %	GPT-1 %
AQuA 80%	49.6 (5.21)	62.4 (0.66)
AQuA 60%	42.8 (13.6)	60.8 (0.50)
AQuA 40%	42.3 (2.23)	57.1 (0.48)
AQuA 20%	39.6 (7.19)	49.5 (0.59)
SWAG+AQuA 80%	64.5 (3.46)	65.3 (0.55)
SWAG+AQuA 60%	67.3 (0.62)	63.6 (0.85)
SWAG+AQuA 40%	64.8 (0.62)	60.6 (0.37)
SWAG+AQuA 20%	60.3 (2.98)	56.3 (0.51)
SWAG only	42.1	38.1
AQuA (Answer only)	28.4	53.9

Table 3: Accuracy of BERT and GPT on different training settings when tested on AQuA. Numbers in parentheses represent the standard deviation.

4.4 Human Evaluation

For each category, we measure the accuracy of the BERT and GPT models trained on SWAG+AQuA. We also measure human accuracy as a baseline. Human accuracy was calculated as the mean accuracy of three human annotators, covering 707 dataset questions in total. Human annotators answered 95.3% of questions correctly, presenting a 7-fold reduction in error compared to the fine-turned BERT model. Inter-annotator agreement was computed over a set of 50 additional questions with a pairwise average Cohen-Kappa score of 0.89, which is interpreted as almost perfect agreement by some guidelines. Table 4 displays the accuracy of the human annotators and neural networks on each category.

5 Discussion

Based on our experiments, we find that model performance on AQuA is substantially lower than those seen on SWAG, which has seen models achieve over 85% accuracy. We observed a decrease of 19.2% on BERT and 14.9% on the OpenAI GPT-1 models between the accuracy on SWAG and the accuracy on our SWAG+AQuA

Question Type	Human %	BERT %	GPT-1 %
Idioms	97.5	69.5 (4.44)	72.6 (1.33)
Reference	100	63.1 (4.08)	71.0 (2.04)
Polysemy	91.7	62.9 (4.93)	55.2 (3.40)
Negation	100	60.0 (5.37)	60.5 (2.14)
Quantitative	97.6	51.5 (1.82)	49.5 (3.80)
Other	94.9	64.9 (4.33)	65.7 (0.54)
Total	95.3	64.5 (3.46)	65.3 (0.55)

Table 4: Class-wise and overall accuracy of human annotators and neural network models, sorted by BERT performance on the proposed categories. Numbers in parentheses represent the standard deviation.

setting. This is especially significant since human error on AQuA is 4.7%—less than a third of the 15% expert error on the SWAG dataset. This suggests that AQuA is challenging to our QA systems because of the difficult commonsense reasoning involved, and not because of ambiguity or intractability in the dataset.

5.1 Question Categories

The logic categories including Quantitative and Negation are especially difficult for our models, seeing some of the lowest accuracies from both models, in contrast to the 99.0% weighted average human accuracy on these categories. Surprisingly, both models performed very well on the Idioms category, suggesting that our neural systems may be capable of learning idioms just like other semantic knowledge. Further identification of additional distinctive and interesting categories that cover the entire dataset may prove very useful in directing our efforts towards aspects of our commonsense QA systems that require the most attention.

5.2 Annotation Artifacts

Annotation artifacts are known to exist in many datasets and may be exploited by supervised models to achieve inflated performances (Gururangan et al., 2018). In AQuA, we did not explicitly filter questions with artifacts or try to detect them. We instead incentivize the question authors, who have some knowledge of how the learners work, to avoid introducing noticeable artifacts in their submissions, as explained in Section 3.1. Our results show that artifacts do not provide sufficient signal for state-of-the-art neural models to come close to human-level accuracy on our data.

5.3 Answer-only Baseline

In the answer-only experiment (where questions are omitted during training and testing), we found that BERT achieves 28.4% accuracy, only slightly above random, whereas GPT-1 achieves 53.9% accuracy, which is the equivalent of narrowing four random options down to two. By comparing this to the AQuA experiment setting, we can interpret these results as an indication of the extent to which the signal was in the answers. While this could be due to artifacts, such as the right answer commonly being of a certain length, we also observed that in many cases, distinguishing between reasonable and ridiculous answers (without seeing the premise) is a part of commonsense reasoning. For example, a commonsense reasoner would be able to rule out the choice "picks up his phone and calls his mom to tell her he doesn't have his phone" without seeing the premise, as a contradiction is contained in the answer. Similarly, "kicks a field goal, celebrates by transforming into a fish, and then quits football" is unlikely to be veracious regardless of the hidden subject.

5.4 Dataset Size

Our experiments show that AQuA forms a challenging extension to the existing SWAG dataset. Even when we train a system to perform near human-level on SWAG, and then fine-tune on AQuA, the system still struggles to answer AQuA questions correctly. However, AQuA is also smaller than SWAG. Our results do *not* suggest that AQuA questions are more difficult than SWAG questions if dataset size is equalized. When we restrict to a subset of SWAG of the same number of questions as AQuA, we find that SWAG has comparable accuracy for GPT (63.6% on reduced-size SWAG vs 62.4% for AQuA) and much lower accuracy for BERT (28.7% vs 49.6%). This shows that AQuA questions are distinct from and complementary to SWAG questions, but taken in isolation are not necessarily more challenging.

Our results suggest two recommendations for dataset construction which we hope to evaluate in future work. The first is, rather than using a single protocol to collect one monolithic dataset, the community may be able to obtain more challenging data by aggregating a variety of distinct, independently-gathered datasets that follow a similar format. For example, pre-training on SWAG and evaluating on AQUA forms a more challenging benchmark than training and testing on SWAG alone. Secondly, if we wish to use our adversarial collection approach to grow AQuA to tens of thousands of examples, we should update our system as new data arrives, so that contributors are able to tune their questions to remain difficult for the strongest, most up-to-date version of the system. Under such a data collection scheme, we may need to increase the reward for fooling the model in cross-validation compared to that for fooling the current model (whereas, these two rewards were equal in AQuA), in order to disincentivize adversarial attacks that manipulate the current model to make it easy to fool on subsequent questions.

Our experiments on different sizes of AQuA produce very different results for BERT and GPT. Unsurprisingly, GPT performance improves with more data on both the AQuA-only and SWAG+AQuA experiments, with the rate of improvement slowing down as data size increases. However, the BERT results are more challenging to interpret. On the AQuA-only setting, BERT appears to improve with data size, but the extremely high variance prevents us from being certain of any trend in BERT's performance on this setting. The variance is lower in the SWAG+AQuA setting and accuracy increases as data size goes from 20% to 60%, but accuracy decreases between SWAG+AQuA-60% and SWAG+AQuA-80% settings (although the SWAG+AQuA-80% setting has high variance and the true mean may be higher). The inconsistency in improvement with more AQuA data after training on SWAG+AQuA-60% for BERT and the reduced rate of performance gain for GPT suggest that it is unclear whether the performance of all models will improve dramatically with an even larger AQuA dataset size.

6 Conclusion

We present AQuA, a commonsense question-answering dataset that is adversarially constructed by allowing humans to view feedback from a pre-trained model and use this information to design challenging commonsense questions. Our experimental results show that AQuA questions present a complementary extension of the SWAG dataset, testing additional modes of common sense.

We identify specific categories of commonsense questions to determine types of reasoning that are more challenging for existing models. In partic-

ular, we note that Quantitative questions have low accuracy for both BERT and GPT. A more detailed analysis into why models struggle to reason about numbers as well as development of more detailed categories of commonsense reasoning are items for future work.

Acknowledgments

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

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*.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. 2017. Towards linguistically generalizable nlp systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10.

Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1460–1469.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

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*.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

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.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1112–1122.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.