Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, pages 1807–1810. ACM.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4201–4207. ijcai.org.

Xiao Liu, Heyan Huang, and Yue Zhang. 2019a. Open domain event extraction using neural latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871, Florence, Italy. 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. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Shangwen Lv, Wanhui Qian, Longtao Huang, Jizhong Han, and Songlin Hu. 2019. Sam-net: Integrating event-level and chain-level attentions to predict what happens next. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6802–6809. AAAI Press.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5):1106–1115.

Fred Morstatter, Aram Galstyan, Gleb Satyukov, Daniel Benjamin, Andrés Abeliuk, Mehrnoosh Mirtaheri, KSM Tozammel Hossain, Pedro A. Szekely, Emilio Ferrara, Akira Matsui, Mark Steyvers, Stephen Bennett, David V. Budescu, Mark Himmelstein, Michael D. Ward, Andreas Beger, Michele Catasta, Rok Sosic, Jure Leskovec, Pavel Atanasov, Regina Joseph, Rajiv Sethi, and Ali E. Abbas. 2019. SAGE: A hybrid geopolitical event forecasting system. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6557–6559. ijcai.org.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. 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.

Ulrich Pilster and Tobias Böhmelt. 2014. Predicting the duration of the syrian insurgency. *Research & Politics*, 1(2):2053168014544586.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. 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 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,

Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Vasanthan Raghavan, Aram Galstyan, and Alexander G. Tartakovsky. 2013. Hidden markov models for the activity profile of terrorist groups. *Ann. Appl. Stat.*, 7(4):2402–2430.

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, Austin, Texas. Association for Computational Linguistics.

Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Paul Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Maria Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014a. 'beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1799–1808. ACM.

Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Paul Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Maria Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014b. 'beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1799–1808. ACM.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Sebastian Schutte. 2017. Regions at risk: Predicting conflict zones in african insurgencies. *Political Science Research and Methods*, 5(3):447–465.

Yawei Sun, Gong Cheng, and Yuzhong Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 806–817, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Tetlock, Barbara A. Mellers, and J. Peter Scoblic. 2017. Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355:481–483.

Philip E Tetlock and Dan Gardner. 2016. *Superforecasting: The art and science of prediction*. Random House.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

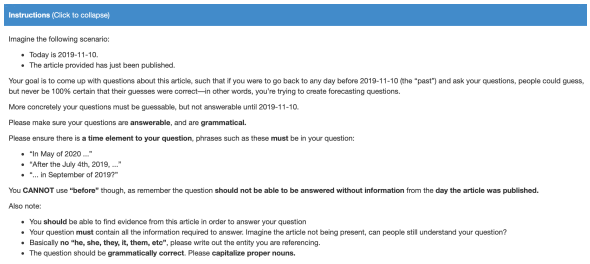

Figure 7: Instruction of creating multiple-choice questions.

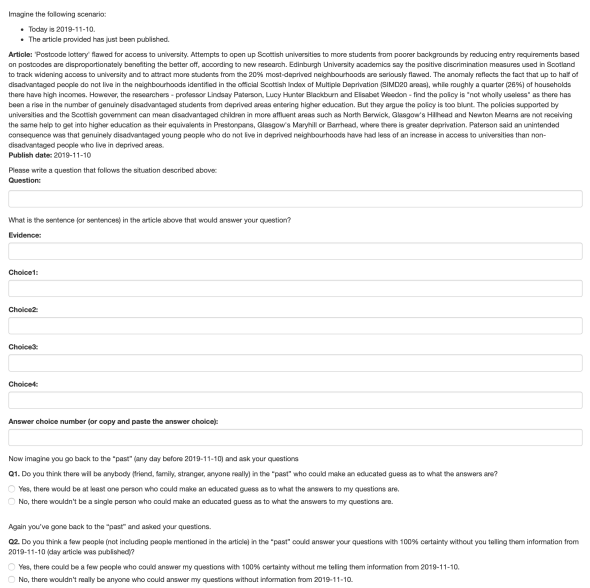

Figure 8: Interface of creating multiple-choice questions.

A Detailed Dataset Creation

In this section, we present detailed explanations of dataset creation. We first selected news sources as in the following section.

A.1 List of News Sources

The New York Post, The New York Times, New York Magazine, Daily News (New York), The Washington Post, NPR All Things Considered, NPR Weekend Edition Saturday, NPR Morning Edition, CNN Wire, CNN.com, CNNMoney.com, CNN INTERNATIONAL, Fox News Network, York Guardian, Washingtonpost.com, The Washington Post Magazine, thetimes.co.uk, Guardian Weekly, Russia & CIS General Newswire, US Official News, The Times (London).

A.2 Dataset Creation

Turking Guidelines. Figs 7 and 8 show the instructions and interface for creating our multiple-choice questions. Workers made multiple-choice distractors with their own knowledge, but they were

Figure 9: Interface of verifying questions.

encouraged to find good distractors using search engines. To ensure the answerability of the created questions, we ask them to indicate the answer along with the supporting evidence that the question is made from. We omit the interfaces due to the space limit.

Initial Screening. The ideal result of our crowdsourcing task are forecasting questions that are tractable but not trivial, and by definition not answerable with certitude using information currently available. Thus to avoid undesirable questions, we asked two additional questions to help screen poorly constructed questions. As shown in Fig 8, we try to determine the difficulty of the question and whether it is answerable using "current" or "past" information. Question 1 attempts to establish whether the question is indeed tractable and asks whether there exists some qualified group of people who could reason and make an educated guess at the answer to the question. On the other hand, question 2 tries to determine if the question is either too easy or is definitively answerable given "current" and "past" information. Thus, the desired response is "yes" and "no" for Questions 1 and 2, respectively; we filtered out created questions that do not satisfy the desired condition.

A.3 Additional Question Quality Checks

We asked the same two questions from our initial quality screening and an additional question to help adjust the timestamp associated with the question if needed. Per question, we got 3 crowd workers to answer the three questions and took the majority vote for question 1 and 2, while selecting the earliest selected timestamp for question 3. We dropped the question, if the majority vote was "no" for question 1 or "yes" for question 2. Moreover, if at least one worker selected "e" in the question 3 (There is no appropriate recent time stamp), then we filtered out the question. Additionally, if the created ques-