EVALUATING THE ROBUSTNESS OF MACHINE READING COMPREHENSION MODELS ON LOW RESOURCE ENTITY RENAMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Question answering (QA) models have shown compelling results in the task of Machine Reading Comprehension (MRC). Recently these systems have proved to perform better than humans on held-out test sets of datasets e.g. SQuAD, but their robustness is not guaranteed. The QA model's brittleness is exposed when evaluated on adversarial generated examples by a performance drop. In this study, we explore the robustness of MRC models to entity renaming with entities from low resource regions such as Africa. We propose EntSwap, a method for test-time perturbations to create a test set whose entities have been renamed. In particular, we rename entities of type: *country, person, nationality, location, organization* and *city* to create AfriSQuAD2. Using this test set we evaluate the robustness of three popular MRC models. We find that compared to base models, large models perform comparatively well on novel entities. Furthermore, our analysis indicate that *person*, as an entity type, highly challenges the model performance.

1 INTRODUCTION

Machine reading comprehension (MRC) is a question-answering task over unstructured text with the aim of examining the understanding and reasoning capability of a model. Over the past few years, there has been growing interest in this task due to the availability of large-scale datasets such as SQuAD (Rajpurkar et al., 2016), MS MARCO (Nguyen et al., 2016). Furthermore, the advent of deep learning techniques and frameworks (Vaswani et al., 2017; Devlin et al., 2019; Sukhbaatar et al., 2015) has improved the performance of MRC models as shown in some model performance on some specific tasks as compared to humans (Rajpurkar et al., 2016).

Despite the impressive performance, these models show poor performance compared to humans on adversarial attacks, given the results in recent works, they are still not robust to all language understanding for adversarial attacks (Belinkov & Bisk, 2018; Jia & Liang, 2017) and out of distribution examples (McCoy et al., 2020; Talmor & Berant, 2019). Several works have proposed investigating the robustness of MRC models to test-time perturbations by creating adversarial examples (Jia & Liang, 2017). One of the earlier works by Jia & Liang (2017) appended semantically irrelevant sentences containing a fake answer that resembles the question syntactically to the context to confuse the model. The authors showed how fragile SQuAD models are with the introduction of out-of-distribution phenomena whereby perturbing the test set yields close to a 50% drop in generalization performance.

The notion of MRC robustness has been investigated in several other different settings (Si et al., 2021; Jin et al., 2020; Liu et al., 2022).Recently Yan et al. (2022) created adversarial examples by renaming entity names in the several datasets with novel entities. The authors proved there is a discrepancy in model performance between answer entities observed during training time and novel answer entities. Our work builds on this idea to investigate the robustness of MRC models in renaming English entities with African-based entities. We leverage the method of entity swapping to create a test dataset. In particular, we investigate the distribution shift at test-time caused by entities from the African region (e.g., country and city) names. Using existing MRC datasets, SQuAD2 (Rajpurkar et al., 2016) we create adversarial examples to evaluate the robustness of span-based MRC models to test-time perturbations.

It is argued that a robust model even though it has observed a small subset of all possible entity names available, should be able to generalize to novel entities. Though simple and understudied, entity swapping tests the model's capability to generalize to novel entities due to a large number of possible entities. In addition, an entity name has world knowledge associated with it and this may change at any given time. Thus, MRC models should not overly rely on specific entities as this would lead to poor generalization of novel entities. Therefore in this study, we first investigate the distribution of entity names from both the train and dev set of SQuAD2. We show that the most common entity names are from high-resourced regions (e.g., Europe, America, etc.) compared to Africa. With such, we investigate the robustness of MRC models in answering questions and extracting answers with entity names from Africa.

Our key contributions are as follows: 1) We propose a method to create adversarial examples with entities from low-resource regions such as Africa. Since most of these regions have fewer digitized articles on Wikipedia, this method can be used to ensure a fair representation of regions during dataset creation. 2) We provide a detailed analysis on the robustness of MRC models to entity names from Africa. We show that although large models generalize comparatively well to novel entities, there is still a performance drop. 3) In our error analysis, we highlight the factors affecting the model's performance, thereby limiting its robustness to entity renaming, which we believe will foster future research towards more robust MRC models.

2 Related Work

The use of adversarial examples to evaluate and improve the robustness of machine learning models has a long-standing history (Holmstrom et al., 1992; Wager et al., 2013). In the field of NLP, one of the early works by Jia & Liang (2017) showed that despite existing neural network QA systems proving their success when evaluated on standard metrics, they perform poorly when evaluated on adversarial examples. In their work, they propose the creation of adversarial examples for SQuAD v1.1 using the AddSent and AddAny algorithms. In AddSent, a distractor sentence is appended at the end of each context. For the AddAny algorithm, a random sequence of grammatical and ungrammatical words are appended to each context. They retrained the BIDAF model (Seo et al., 2017) on these generated adversarial examples to test its robustness. AddSent algorithm swaps named entities and numbers in the question with the nearest word in GloVe word vector space (Manning et al., 2014). Our method, EntSwap, slightly differs from AddSent. We replace named entities but left numbers unchanged. Unlike AddSent, which replaces entities in the question, appends distracting sentences to the context, and leaves the answers unchanged, EntSwap replaces all detected named entities for the questions, context, and answers to create an altered SQuAD2 dev set.

Although the BIDAF model (Seo et al., 2017) was retrained on adversarial examples, there is no guarantee of its robustness when evaluated on differently generated adversarial examples. Wang and Bansal (Wang & Bansal, 2018) generated slightly different adversarial examples for SQuAD using the AddSentMod algorithm by prepending the distractor sentences to the context instead of appending them and also used a different set of fake answers from ADD-SENT. The authors showed that the pre-trained BiDAF model (Seo et al., 2017) was not robust to this set of adversarial examples as the model's F1-score dropped by 30%. While both works use distractor sentences to create adversarial examples, our study randomly swaps original English named-entities with entities from Africa.

Similar to our work, Yan et al. (2022), create adversarial examples by renaming entity names in several datasets, including SQuAD, with novel entities. Their perturbation involved swapping the answer entity in the whole passage for the categories, person names, organizations, and geopolitical entities. Unlike (Yan et al., 2022), our work is different in the following ways: first, we swap six categories of entity type i.e., person, city, country, organization, nationality, and location. Secondly, we swap all entities detected in the context for the six categories. Every entity swapped in the context is swapped in the question, answer, and article title.

3 METHODS

3.1 DATA

We leverage the existing extractive MRC dataset, SQuAD2.0 (Rajpurkar et al., 2018), where we apply our perturbation method on the dev set to create a dev set that we named AfriSQuAD2. We choose to conduct our evaluation on SQuAD2.0 because 1) SQuAD2.0 format is extractive QA, i.e the answers are short and are a span from the passage, and 2) Questions or answers are named-entities. This allows us to test the model's capability to answer a question on a novel entity or extract a novel entity as an answer.

3.2 MRC MODELS

We experiment with three pretrained language models, which have shown comparative performance on the popular SQuAD2.0 benchmark with base and large variations of the models. **BERT** (Devlin et al., 2019), a pretrained deep directional encoder trained on English Wikipedia and BookCorpus, with masked language modeling (MLM) and next sentence prediction (NSP) as the pretrained objectives. Unlike BERT, **RoBERTa** (Liu et al., 2019) has shown better performance using only the MLM training objective and pretraining on a large diverse corpus. Compared to BERT and RoBERTa, **DeBERTa** (He et al., 2021) uses a much larger model size and training corpus, which allows it to capture more complex relationships between words and sentences in a language.

For model evaluation, we use the models finetuned on the original SQuAD2 training data from deepset [1] publicly available on Huggingface (Wolf et al., 2020). For BERT and RoBERTa, we use the uncased and distilled versions respectively.

4 SWAPPING ENTITY NAMES

In this section, we describe EntSwap, our method for perturbing a MRC dev set, by renaming entities with named-entities of African origin. We also describe how we generate collections with entity names for substituting the six categories, with the aim to study the model's robustness.

4.1 PERTUBATION METHOD

Adversarial examples have been created using different techniques such as perturbing by text paraphrasing (Iyyer et al., 2018), character-level typos insertion (Belinkov & Bisk, 2018), appending to the input distracting Jia & Liang (2017), and replacing occurrence of certain words in the text with corresponding words (Alzantot et al., 2018). In our work, we create SQuAD2.0 adversarial examples by randomly swapping the entities. We use the EntSwap algorithm to create an altered SQuAD2.0 dev set with a large percentage of entities from Africa. Figure 1, show our perturbation method which we introduce below.

Step1: Named Entity Recognition. In order to identify entities to swap, we run a named-entity recognizer (NER) with the Stanza version of the Stanford CoreNLP tool (Qi et al., 2020) on the context, questions, answers, and titles. NER has proven to be a core component in question-answering tasks, especially extractive question-answering. We use the Stanford CoreNLP tool because of its ability to identify individual GPE type entities (i.e., county and city) unlike other tools which have them under the GPE tag. In this work, we identify six entity types: *Person* , *Country* , *City* , *Location* , *Organization* and *Nationality* . We mainly focus on these entity types because of their frequent appearance on the question or answer and have the large possibility of valid names.

Step2: Identifying span. Answers in the SQuAD2.0 dataset have the start and end position of the answer span. Borrowing from this, to identify the span of the entity name to be swapped, during NER we assign the start and end position of the identified entity. With the start and end position, we ensure the entity is swapped with another entity of the same length. The number of perturbable spans is shown in Table 1, for the train and dev set. We only swap entities in the dev set.

[1] https://www.deepset.ai/

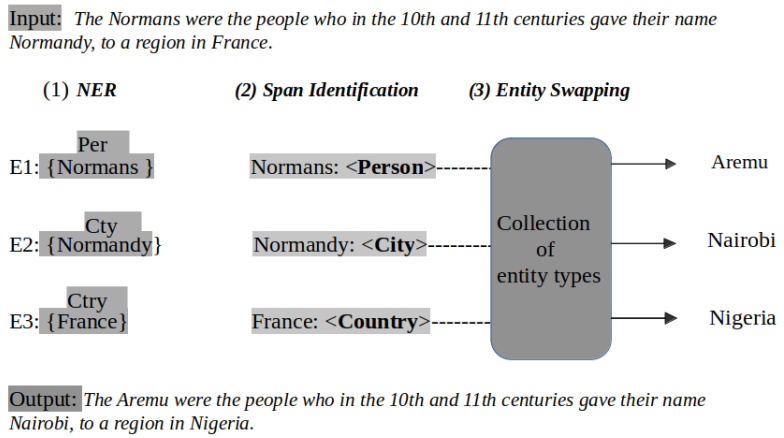

Figure 1: The perturbation method for swapping entity name in an MRC dataset

category	train	dev
person	50706	2563
organization	27550	2041
location	22327	1581
city	15529	1114
country	20633	1184
nationality	16792	1104

Table 1: Distribution of named entities by categories in the train and dev set of the experimental dataset

Step3: Entity sampling and swapping. After the NER tool has identified the entity span to be swapped, we obtain the entity name to be swapped from a collection of entities of the same entity type. The entity name is randomly sampled from a set of entity names from the collection of the same entity type.

Given a candidate entity name for each pertubable span, we do a string match on the context, question, answer, and title. If an entity occurs more than once in the same context, it is replaced by the same entity name in all scenarios. An entity name appearing in more than one context may or may not be replaced with the same entity name. To cater for entities that are inflections of another entity, we string match the main entity and maintain the inflection. For example, *normans* is an inflection of *norman*, to maintain this we substitute *norman* with *Aremu* and *normans* with *Aremus*.

For entity type person, to ensure most of the names selected are of African origin, we select the second name or second and third names of a person. This is because most first names are English names and to ensure we evaluate the models on novel entities, we choose not to use these names.

4.2 Collection of Entity Names

Using the pre-defined categories in subsection 4.1, we scrape named entities from the Wikidata knowledge graph (Vrandecic & Krötzsch, 2014) of African origin. The collections are created by extracting canonical names (e.g. Kenya, Abidemi, Tripoli) of six different named-entities. We use SPARQL queries to search Wikidata for named entities of each category type using relations such as:

'COUNTRIES IN AFRICA', *'CITY OF A COUNTRY, A COUNTRY IN AFRICA'*, and *'PERSON BORN IN A CITY, A CITY IN A COUNTRY, A COUNTRY IN AFRICA'*.

We further ensure there are no duplicate entries by getting rid of duplicate Qids and all entities represented with Qids are also manually deleted.

Each entity type is dumped into a CSV file of it's own. In total, we create six different collections: person, country, nationality, city, and organization with entities from Africa.

4.3 Entity Swapping Quality

The performance of the MRC models is dependent on the quality of our perturbation method. We therefore randomly sample 50 contexts from the dev set and manually check for the quality of the perturbed spans. We evaluate the accuracy of step2 and 3, that is identifying the perturbable span and swapping the span with the novel entity for categories *Person, Country* and *City*. Results reported in Table 2, show that our method gets acceptable accuracy, thus confirming the quality of the perturbed example.

	Accuracy		
Step	PERSON	COUNTRY	CITY
Span identification	94.32	92.17	91.26
Entity Swapping	88.85	96.37	94.61

Table 2: Accuracy of the two key steps in our perturbation method on 50 randomly sampled contexts from AfriSQuAD2. We report the percentage accuracy.

5 Results and Analysis

In this section we report the results of our experiments and an in depth analysis to understand which entity types pose a great challenge to MRC models' robustness. We report our results based on the F1-score ans exact match (EM) metrics.

5.1 How representative are the entity names in MRC datasets?

To understand how representative the entity names in SQuAD dataset an MRC datset are, we analyze the entity types *city* and *country*. We get the most frequent 14 entities in the train ant dev set and map them to their relations on wikidata. Especially their geo-political representation. Figure 2 shows the top 14 entities in the train and dev set for entity type city in Figures 2a and 2b respectively, and entity type country in Figures 2c and 2d. We note that 90% of the entity names are from either Europe or American continent, thus showing how most articles in the SQuAD2 and other MRC datasets in general are not representative of low resource regions. most MRC datasets rely on Wikipedia as a source for their dataset creation and this is influenced by the number of articles available on particular regions. In addition, low-resource regions often have a lower proportion of digital resources, meaning that there is less text data available for training MRC models. This can result in less robust models and datasets that do not reflect how diverse the world is.

5.2 How robust are MRC models to entity renaming?

We conduct evaluation on the original validation set of SQuAD2 and on the perturbed AfriSQUAD2 datasets.

We report the performance of several models on AfriSQuAD2 and SQuAD2 dev set in Tables 3 and 4. From the results, we note that: 1) All models show performance drop on AfriSQuAD2 as compared to the original dev set of SQuAD2. 2) BERT-base is the most vulnerable model to AfriSQuAD2 with respect to both metrics. This indicates that models that have high performance on the original dev set generally tend to perform better also on adversarial attacks. 3) Large models suffer less from adversarial attacks compared to base models. This is due to their large size and capacity. Thus, they have the ability to capture more complex patterns and relationships in data, such as identifying an answer span with a novel entity name and in turn improving their accuracy.

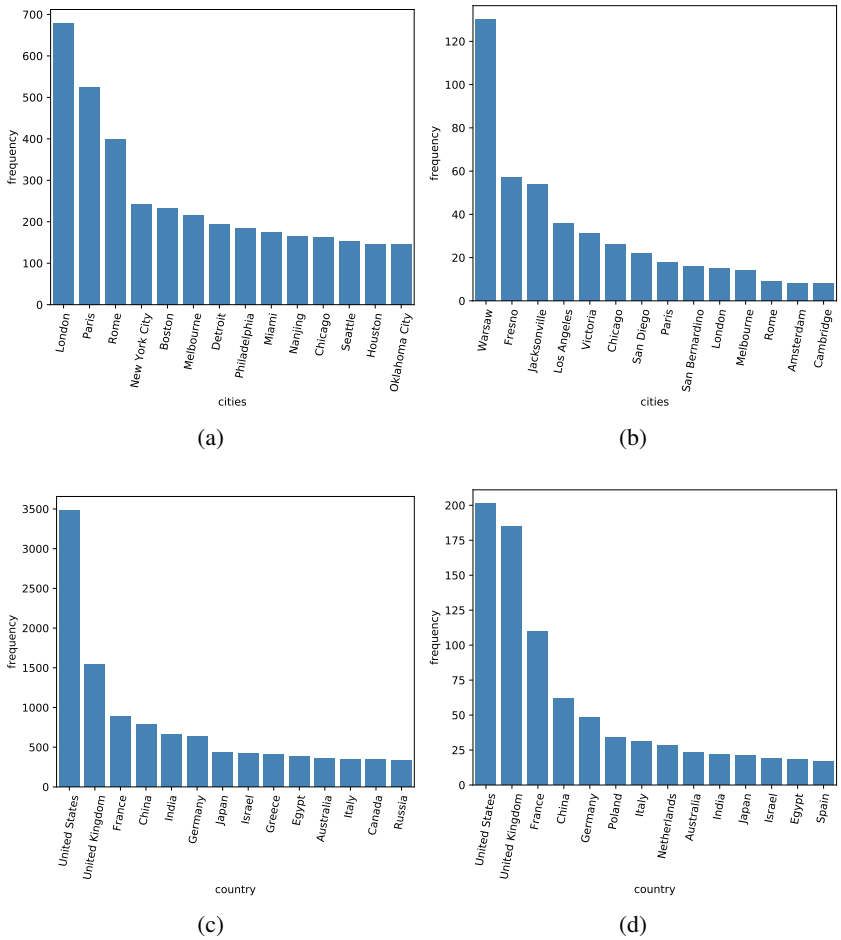

Figure 2: Distribution of top 14 named entities occuring for city category in (a) train set and (b) dev set, and for country category in (c) train set and (d) dev set.

	DeBERTa-large		RoBERTa-large		BERT-large	
Dataset	EM	F1	EM	F1	EM	F1
SQuAD2	88.07	91.14	85.08	88.26	80.83	83.83
AfriSQuAD2	84.14	87.54	81.60	84.96	79.29	82.52
△	**3.93**	**3.60**	3.48	3.30	1.54	1.31

Table 3: Performance comparison of several MRC models on the SQuAD2 and AfriSQuAD2 dev set. We report the F1-score and exact match for both datasets. We represent the difference between the performance of the models on the two datasets with △. We boldface models with the greatest performance drop.

5.3 Which entity types are more challenging to MRC Models

In Table 5, we present the performance of the BERT-large model on a different combination of entity types. AfriSQuAD2-w/o-L, where L represents the entity type, means we replaced all the other five entity types except entity type L. For example, AfriSQuAD2-w/o-city means we swapped all other entity types except entity type city. Per, Org and Loc renaming prove to be challenging to the model's robustness performance. We note that when we do not have these entity types in the AfriSQuAD2 dev set the model performs as close as the AfriSQuAD2 dataset. This indicates that renaming these entity types poses a challenge to the robustness of MRC models. This is likely because for *person*

Dataset	DeBERTa	RoBERTa	BERT
SQuAD2	88.07/83.84	85.08/80.35	80.83/75.57
AfriSQuAD2	84.14/80.32	81.60/78.05	79.29/74.42

Table 4: Performance comparison of different MRC model sizes. EM scores for the *LARGE/BASE* variants of models on SQuAD2 and AfriSQuAD2 .

Dataset	EM	F1
SQuAD2	80.83	83.83
AfriSQuAD2	79.29	82.52
AfriSQuAD2 w/o city	78.54	81.90
AfriSQuAD2 w/o country	79.57	82.92
AfriSQuAD2 w/o location	80.17	83.39
AfriSQuAD2 w/o nationality	79.68	82.99
AfriSQuAD2 w/o organization	80.61	83.68
AfriSQuAD2 w/o person	80.08	83.29

Table 5: Comparison of BERT-large performance on different entity types in the AfriSQuAD2 dev set.

names, we only swapped from the second name and avoided the first name, implying that most of the names are of African origin, which the models may not have been exposed to during training. This also applies to Organization and Location entity types, most of the organizations in the collection are small, local organizations within individual countries, which the model may have not access to during training. In the train and dev set of SQuAD2, we note that these entity types are the most frequent in the dataset with 50k Person names, 27k Organization names, and 22k Location names. This implies that many of these entities are swapped in the dataset compared to CITY, COUNTRY, and NATIONALITY entity types.

5.4 Analysis

We do an analysis based on the BERT model to understand the model's behavior towards AfriSQuAD2 examples. We focus on examples where the models originally predicted the correct exact match but failed on the altered examples. We also focus on cases where the model had a low confidence score compare to the original data set, even when the model extract the correct answer span.

Error analysis. Table6 shows the EM score of BERT-large on HasAns and NoAns questions. Compared to SQuAD2, the performance of BERT on AfriSQuAD2 for HasAns is low in comparison to NoAns (15.56% drop). This shows that the model is able to predict with high accuracy when a question is unanswerable, unlike a question with answers. Hence, we seek to understand why the performance of BERT drops on HasAns questions on AfriSQuAD2. We randomly sample 100 questions and classify them into HasAns (56%) and NoAns (44%) based on the ground truth. We note that 40% of the HasAns questions were wrongly predicted as NoAns questions. This mostly happened when either the question was about a novel entity or the answer to be extracted was about a novel entity. In particular for a person entity type, when the two names are of African origin, they have a huge influence on the model's performance.

Dataset	HasAns	NoAns
SQuAD2	79.41	82.25
AfriSQuAD2	71.50	87.06

Table 6: EM score of BERT-large on questions with answers (HasAns) and unaswerable questions (NoAns).

Model success and failure. Language models such as BERT are pre-trained on large unstructured text corpora with diverse named entities. They leverage being exposed to a diverse set of named entities. This is why the performance BERT on AfriSQuAD2 does not drop drastically. We cannot guarantee the same performance for non-transformer based models, especially those not trained on large corpora with diverse named entities.

Although we report over 80% accuracy in detecting the span and swapping an entity in the dev set, we can not quantify the role of data quality in the model performance. and thus some of the performance drops may be attributed to data quality.

6 CONCLUSION

In this work, we study the robustness of MRC models to entity name swapping. In particular, we propose EntSwap, a method for swapping entity names from the original dataset along with collections of entities of six different categories of African origin. With this method, we create AfriSQuAD2, by renaming entity names in the original SQuAD2 dev set with ones from our collections. We experiment with three popular MRC models using SQuAD2 dev set and AfriSQuAD2. Though the original SQuAD format was maintained, we find that the AfriSQuAD2 examples challenge the capability of MRC models to extract the correct answer span when the question is about a novel entity. Swapping of entity types: Person, Organization, and location poses the greatest challenge to MRC models. The drop in performance of the models can be attributed to the over reliance of MRC models on real-world entity knowledge. In addition, we observe that most entity names are from high-resource regions,thus, these models may have not been exposed to a subsample of entities in our collections.

For future work, we would like to extend this study to more datasets created from different sources. For example, we would like to see how these models will perform on a perturbed test set of a dataset created via distant supervision compared to a human-created dataset like SQuAD2.

REFERENCES

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1316. URL https://www.aclweb.org/anthology/D18-1316.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Lasse Holmstrom, Petri Koistinen, et al. Using additive noise in back-propagation training. *IEEE transactions on neural networks*, 3(1):24–38, 1992.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1875–1885, New Orleans, Louisiana,

June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1170. URL https://www.aclweb.org/anthology/N18-1170.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://www.aclweb.org/anthology/D17-1215.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8018–8025. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6311.

Linqing Liu, Patrick S. H. Lewis, Sebastian Riedel, and Pontus Stenetorp. Challenges in generalization in open domain question answering. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 2014–2029. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-naacl.155. URL https://doi.org/10.18653/v1/2022.findings-naacl.155.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The {Stanford} {CoreNLP} Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupala, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pp. 217–227. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.21. URL https://doi.org/10.18653/v1/2020.blackboxnlp-1.21.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (eds.), *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In Asli Celikyilmaz and Tsung-Hsien Wen (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pp. 101–108. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-demos.14. URL https://doi.org/10.18653/v1/2020.acl-demos.14.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://www.aclweb.org/anthology/P18-2124.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ0UKP9ge.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. Benchmarking robustness of machine reading comprehension models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 634–644. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.56. URL https://doi.org/10.18653/v1/2021.findings-acl.56.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 2440–2448. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Alon Talmor and Jonathan Berant. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4911–4921. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1485. URL https://doi.org/10.18653/v1/p19-1485.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 351–359, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/38db3aed920cf82ab059bfccbd02be6a-Abstract.html.

Yicheng Wang and Mohit Bansal. Robust machine comprehension models via adversarial training. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pp. 575–581. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-2091. URL https://doi.org/10.18653/v1/n18-2091.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. On the robustness of reading comprehension models to entity renaming. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 508–520. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.37. URL https://doi.org/10.18653/v1/2022.naacl-main.37.