Factual Consistency of Multilingual Pretrained Language Models

Constanza Fierro   Anders Søgaard
University of Copenhagen
`{c.fierro,soegaard}@di.ku.dk,`

Abstract

Pretrained language models can be queried for factual knowledge, with potential applications in knowledge base acquisition and tasks that require inference. However, for that, we need to know how reliable this knowledge is, and recent work has shown that monolingual English language models lack *consistency* when predicting factual knowledge, that is, they fill-in-the-blank differently for paraphrases describing the same fact. In this paper, we extend the analysis of consistency to a *multilingual* setting. We introduce a resource, MPARAREL[1], and investigate (i) whether multilingual language models such as mBERT and XLM-R are more consistent than their monolingual counterparts; and (ii) if such models are equally consistent across languages. We find that mBERT is as inconsistent as English BERT in English paraphrases, but that both mBERT and XLM-R exhibit a high degree of inconsistency in English and even more so for all the other 45 languages.

1 Introduction

Pretrained Language Models (PLMs) enable high-quality sentence and document representations (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Raffel et al., 2020) and encode world knowledge that can be useful for downstream tasks, e.g. closed-book QA (Roberts et al., 2020), and commonsense reasoning (Zellers et al., 2019; Talmor et al., 2019), to name a few. Recent work has used language models as knowledge bases (Petroni et al., 2019; Kassner et al., 2021a; Roberts et al., 2020) and as the basis of neural databases (Thorne et al., 2021). Such usage of PLMs relies on the assumption that we can generally trust the world knowledge that is induced from these models.

Consistency is a core quality that we would like models to have when we use their stored factual knowledge. We want models to behave consistently

on semantically equivalent inputs (Elazar et al., 2021), and to be consistent in their believes (Kassner et al., 2021b). Moreover we want them to be fair across languages or in other words to exhibit a consistent behaviour across languages (Choudhury and Deshpande, 2021). Nonetheless, recent work on consistency in PLMs has shown that models are brittle in their predictions when faced to irrelevant changes in the input (Gan and Ng, 2019; Ribeiro et al., 2020; Elazar et al., 2021; Ravichander et al., 2020). These works only considered English PLMs, while Jang et al. (2021) studied the consistency of Korean PLMs. There are, to the best of our knowledge, no resources available to measure the consistency of multilingual PLMs.

Contributions   In this paper, we present MPARAREL, a multilingual version of the PARAREL🤘 dataset (Elazar et al., 2021), which we construct by automatically translating the English data to 45 languages and performing a human review of 11 of these. We then evaluate how consistent mBERT is in comparison to its monolingual counterpart, and we study how the consistency of mBERT and XLM-R varies across different languages. Following previous work, we do this by querying the model with cloze-style paraphrases, e.g. "Albert Einstein was born in [MASK]" and "Albert Einstein is originally from [MASK]". We find that mBERT and XLM-R exhibit competitive consistency to English BERT, but consistency numbers are considerably lower for other languages. In other words, while consistency is a serious problem in PLMs for English (Elazar et al., 2021), it is a *much* bigger problem for other languages.

2 Probing Consistency

We use the same probing framework as defined by Petroni et al. (2019) and refined by Elazar et al. (2021), and query PLMs with cloze-test statements created from subject-relation-object Wiki-

---

[1] https://github.com/coastalcph/mpararel

data triples (Elsahar et al., 2018). That is, we have a set of different relations $\{r\}$, and each r has a set of templates or patterns $\{t\}$ and a set of subject-object tuples $\{(s, o)\}$. Each template t describes its corresponding relation r between the pairs (s, o). E.g. a relation r can be *born-in*, and two patterns could be $\{t_1 =$"[X] was born in [Y]", $t_2 =$"[X] is originally from [Y]"$\}$ (where [X] is the subject and [Y] the object to be replaced). Then the corresponding subject-object tuples $\{(s, o)\}$ are used to query and evaluate the model by replacing the subject and masking the object. We study the consistency of a PLM by querying it with cloze-test paraphrases and measuring how many of the predictions of the paraphrases are the same (details in §4).

3   MPARAREL

We used the paraphrases in the PARAREL🤘 dataset (Elazar et al., 2021), which has 38 relations in total and an average of 8.6 English templates per relation. We translated these using the procedure below, obtaining paraphrases for 46 languages.

Translations   We relied on five different machine translation models: Google Translate[2], Microsoft Translator[3], a pretrained mBART model that translates between 50 languages (Tang et al., 2020), a pretrained mixture of Transformers that translates between 100 languages (Fan et al., 2021), and OPUS-MT (Tiedemann and Thottingal, 2020). We fed models with templates, e.g.,"[X] died in [Y]"[4], automatically checking if the translation contained [X] and [Y]. We considered as valid: (1) translated paraphrases that were agreed upon by two or more different models, and (2) the translations from the Microsoft translator, as they were found to be of good quality in several languages as per manual inspection by native speakers. So for languages that Microsoft supports, we will have a template t from the Microsoft translator, as well as any other translation agreed upon by two or more other translators[5]. Finally, we also include the templates in the mLAMA dataset (Kassner et al., 2021a). Translations of subject-object entities were obtained from WikiData, using the entity identifiers. We kept only the languages that (i) covered at least 60% of the

	MPARAREL
Average #relations	37.13
Average total #patterns	343
Min. patterns in a relation	2
Max. patterns in a relation	33
Average patterns in a relation	9.2
Average string distance	13.9

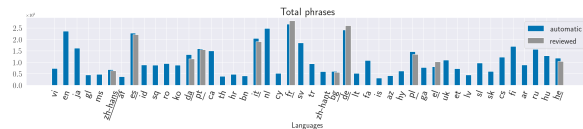Table 1: MPARAREL statistics across languages.

Figure 1: Number of examples per language. Manually reviewed languages are underlined. The order is given by the consistency results (see Figure 2).

total 38 relations,[6] and (ii) covered at least 20% of the total original phrases in English.[7]

Human Evaluation   For assessing the quality of the translated paraphrases we carried out a human review. We had 14 native speakers review 11 different languages[8]. Each person reviewed a 50% random sample of the total templates of the language[9]. We asked whether the template was a correct paraphrase of the given relation, we requested corrections and optionally asked for new template suggestions. On average, $16\%\pm8\%$ of the reviewed templates were considered wrong, $20\%\pm10\%$ were amended, and the rest were considered correct. The statistics of the dataset after removing the wrong templates and including the corrections and suggestions can be found in Table 1. The total number of different phrases (templates with the subject and object replaced) per language is shown in Figure 1.

4   Experiments

We ran experiments with mBERT (Devlin et al., 2019), a multilingual BERT model of 110M parameters trained on 104 languages using Wikipidea, and XLM-RoBERTa (Conneau et al., 2020), a multilingual RoBERTa model of 560M parameters trained on 100 languages using 2.5TB of Common-Crawl data.

---

[2]https://pypi.org/project/googletrans/
[3]https://docs.microsoft.com/en-us/azure/cognitive-services/translator/
[4]Translating populated templates made alignment hard.
[5]In the final dataset, 60% of the templates are agreed by 2 or more translators

[6]Only relations with more than one template with subject-object tuples were included.
[7]A phrase is a populated template.
[8]There were 2 reviewers in Greek, German, and Spanish.
[9]The review took 50 minutes on average and the reviewers did it voluntarily.

Querying Language Models The prediction of a PLM for a cloze statement t is normally $\arg\max_{w \in V}(w|t)$ (Petroni et al., 2019; Ravichander et al., 2020), that is, the top-1 token prediction over the vocabulary. However, Kassner et al. (2021a); Elazar et al. (2021) used typed queries, where the prediction is $\arg\max_{w \in C}(w|t)$, with C a set of candidates that meets the type criteria of the pattern (e.g. cities, professions). In our case, C is all the possible objects in the relation. The motivation is that by restricting the output we can reduce the errors due to surface fluency, as when populating the template with different tuples small grammatical errors can occur (Kassner et al., 2021a).

It is common to only consider tuples (subject-object) for which the to-be-masked object is a single token in the models vocabulary (Petroni et al., 2019; Elazar et al., 2021). However, this reduces the number of valid tuples severely, and even more so when dealing with multilingual vocabularies. Therefore, we follow the multi-token prediction approach in Kassner et al. (2021a) and query the model with multiple masked tokens. The probability of an object instantiation is then the average probability of its tokens, i.e., for a given object $o = w_1 w_2 ... w_l$, $p(o|t) = \frac{1}{l}\sum_{i=1}^{l} p(m_i = w_i | t_l)$, where w_i is the i-th token of the word o, m_i is the i-th mask token, and t_l is the template with l mask tokens.

Evaluation For a given relation r the **consistency** is the percentage of pairs of templates that have the same prediction for every subject-object tuple (Elazar et al., 2021), i.e. the consistency of a given relation r is:

$$\frac{1}{|D|}\sum_{d \in D}\frac{2}{|T|(|T|-1)}\sum_{i=0}^{|T|}\sum_{j=i+1}^{|T|}\mathbb{1}_{f(t_i^d)=f(t_j^d)} \quad (1)$$

where t is a template, T the set of templates in the relation, d is a subject-object tuple, D the set of all tuples, so t_i^d is the i-th template populated with the subject-object data d, and $f(\cdot)$ is the prediction of the model. Next, **accuracy** measures the factual correctness of the predictions and is defined as the percentage of correct predictions over all the templates and data, i.e. $\sum_{d \in D}\sum_{t \in T}\mathbb{1}_{f(t^d)=o}$, where o is the object of the tuple d. Finally, **consistency-accuracy** is the subset of the accurate predictions that is also consistent. Thus, it is computed similarly to Equation 1 but in the indicator's condition we also add the condition imposed in the accuracy.

Metric		BERT	mBERT		
		en	en	ja	zh-hans
Consistency	w/ .	**0.57**	**0.54**	**0.55**	0.46
	w/o .	0.53	0.53	0.52	**0.51**
Accuracy	w/ .	**0.39**	**0.37**	0.13	0.22
	w/o .	0.32	0.35	**0.15**	**0.27**
Consistency-acc	w/ .	**0.32**	**0.3**	0.09	0.15
	w/o .	0.24	0.28	**0.1**	**0.2**

Table 2: Performance comparison of BERT to mBERT, as well as of removing sentence-final punctuation in our input examples, with mBERT results on English, Japanese, and Chinese Simplified.

This metric is useful to account for trivial cases of consistency: A model can be really bad in a language and predict the same token despite the input, and thus be perfectly consistent. For all metrics, we report the *macro average* across relations.[10]

5 Results and Discussion

Table 2 compares the consistency of BERT and mBERT on English data, showing little to no difference, depending on whether we use sentence-final punctuation or not. Sentence-final punctuation is not fully consistent in the machine translation output, so we ran experiments comparing the performance of including sentence-final punctuation or removing it. Since languages vary in how they use punctuation, and sentence-final punctuation causes variance in consistency (e.g., Japanese +3%, but Chinese Simplified -5%), we decided to remove all sentence-final punctuation for the cross-lingual consistency results.

Consistency across languages The consistency results in the MPARAREL dataset are presented in Figure 2. First of all, we can see that the manual corrections don't change the results much (as also experienced by Kassner et al. (2021a)). Nevertheless, they do improve the consistency and accuracy by 1%-2% in a couple of languages, probably because some noise was reduced when correcting and adding new templates. Consistency numbers remain very low, however, especially for other languages than English and Vietnamese. XLM-R is much more consistent than mBERT in *some* languages (e.g. Greek ('el')), yet their average consistency is the same (0.43). The standard deviation of XLM-R's consistency is 8% lower than that of mBERT, i.e., XLM-R has a more fairly

---

[10] Our results are *not* directly comparable to those reported in Elazar et al. (2021), even if we use the same metric, since we filter tuples with the same subject, but two different objects.
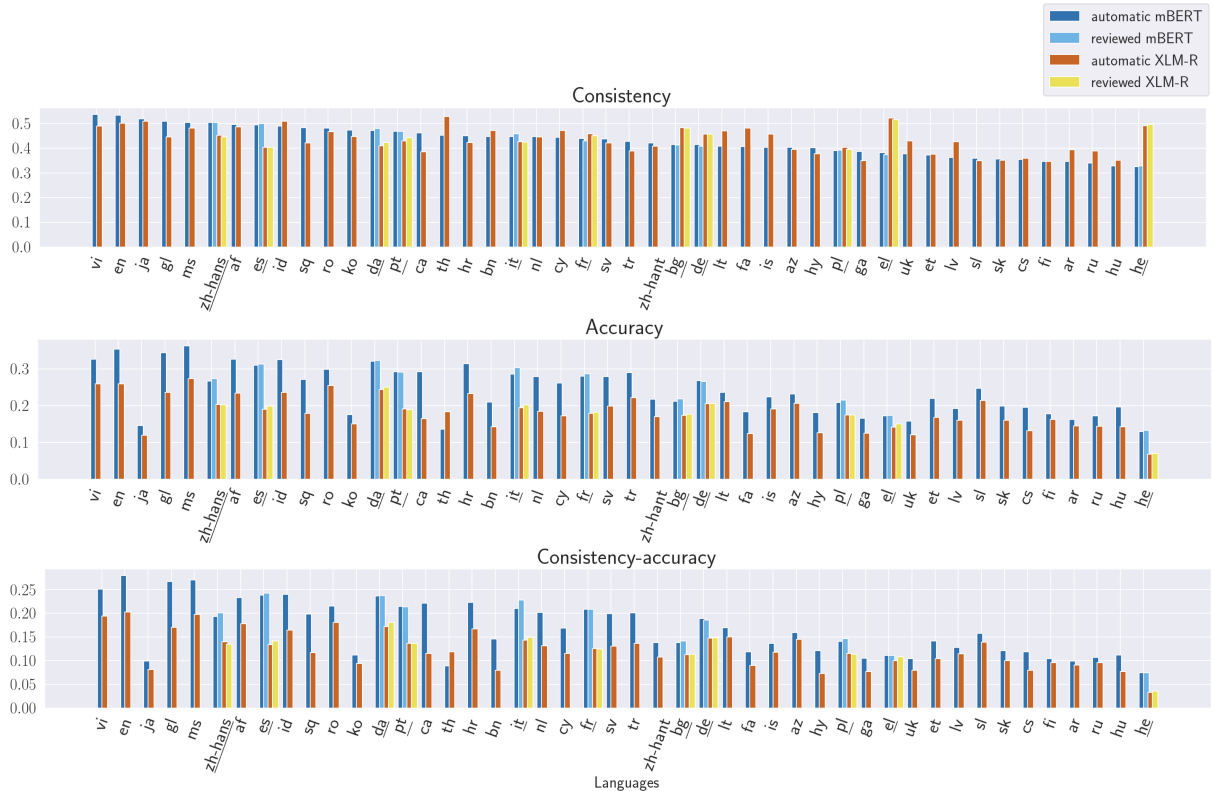
Figure 2: mBERT and XLM-R results on MPARAREL after and before the human review (§3). The order of the languages follows the consistency results in mBERT, and the languages underlined were manually reviewed.

distributed consistency. Somewhat surprisingly, the accuracy of mBERT is superior to XLM-R's, nevertheless, this aligns to the findings of Elazar et al. (2021) where English base BERT obtained higher accuracy than a large English RoBERTa model. We note the importance of controlling for accuracy in our consistency results (reported as *consistency-accuracy*): Japanese, for example, has high consistency, but in part, because it wrongly predicts the same (frequent) token across paraphrases; consistency-accuracy reranks Japanese as one of the most inconsistently encoded languages in both mBERT and XLM-R.

6 Related Work

Petroni et al. (2019); Davison et al. (2019) first studied to what extent PLMs store factual and commonsense knowledge, proposing the LAMA probe and dataset. Then further analysis followed it, Kassner and Schütze (2020) studied probing PLMs factual knowledge on negated sentences, Shin et al. (2020); Reynolds and McDonell (2021); Jiang et al. (2020b) optimized the prompts so to improve the knowledge retrieval, and Bouraoui et al. (2020); Heinzerling and Inui (2021) explored other

approaches different than the cloze-test probing. Then, Kassner et al. (2021a); Jiang et al. (2020a) studied the knowledge memorized in multilingual PLMs, presenting the mLAMA dataset which is a translated version of LAMA.

Consistency in PLMs has been studied in English. Gan and Ng (2019) created a paraphrased version of SQuAD and showed that the state-of-the-art models had a significant decrease in performance, Ribeiro et al. (2020) proposed a framework to test the robustness in the predictions when faced with irrelevant changes in the input. Elazar et al. (2021); Ravichander et al. (2020) showed that monolingual English PLMs are inconsistent in fill-in-the-blank phrases. Then, Newman et al. (2021) proposed using adapters to better handle this inconsistency.

There are paraphrase datasets available in English (Dolan and Brockett, 2005; Quora, 2012) and in multiple languages (Ganitkevitch and Callison-Burch, 2014), but they cannot be easily linked to subject-object tuples in order to measure consistency.

7 Conclusion

In this work, we measured the consistency of *multilingual* Pretrained Language Models when queried to extract factual knowledge. We constructed a high-quality multilingual dataset containing 46 different languages, to assess the consistency of models predictions in the face of language variability. Finally, we experimented with mBERT and XLM-R and concluded that their consistency is poor in English, but even worse in other languages.

References

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7456–7463.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12710–12718.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China. Association for Computational Linguistics.

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

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Ed Hovy, Hinrich Schutze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *ArXiv*, abs/2102.01017.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland. European Language Resources Association.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2021. Accurate, yet inconsistent? consistency analysis on language understanding models. *arXiv preprint arXiv:2108.06665*.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021a. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

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

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021b. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Newman, Prafulla Kumar Choubey, and Nazneen Rajani. 2021. P-adapters: Robustly extracting factual information from language models with diverse prompts. *arXiv preprint arXiv:2110.07280*.

Matthew E. 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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Quora. 2012. Quora question pairs dataset. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs. [Online; accessed 10-November-2021].

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(140):1–67.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm*. Association for Computing Machinery, New York, NY, USA.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. 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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. From natural language processing to neural databases. *Proc. VLDB Endow.*, 14(6):1033–1039.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.