Commonsense Reasoning: how do Neuro-Symbolic and Neuro-only approaches compare?

Ruben Branco[1], António Branco[1], João Silva[1] and João Rodrigues[1]

[1]University of Lisbon
NLX – Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal

Abstract

The representation of knowledge is a central task in Artificial Intelligence and has been an active topic of research since the beginnings of the field. Intensive research and labor has been put into producing resources which encode knowledge regarding different topics, structured in suitable formats so as to allow robust, automated reasoning over them.

In Natural Language Processing, deep learning models are commonly given unstructured data and seek to learn the necessary knowledge and abstractions required to represent and understand the underlying mechanisms that govern the target language processing tasks. A popular method to address this issue is to expand the training process to include more tasks and data. Yet, it remains one of the challenges of deep learning.

In this respect, a promising research path is to combine the rich knowledge encoded in structured resources with deep learning methods, enhancing them with the necessary means to more effectively learn the complexities of the target tasks.

In this paper we set out to compare a Neuro-Symbolic model with mainstream Neuro-only models when they are tasked with solving commonsense reasoning problems, which heavily rely on appropriately represented knowledge: commonsense reasoning is an essential part of the human experience, encompassing human values and needs, and by resorting to it, we can organize sensible arguments and decide on effective actions.

The results obtained indicate that there is no clear advantage to either approach, with the Neuro-Symbolic model being competitive amongst the Neuro-only models, but not superior.

Keywords

commonsense reasoning, neuro-symbolism, transformer

1. Introduction

Given the challenges of current deep learning approaches to Natural Language Processing (NLP) and the availability of rich structured knowledge sources that have been developed and matured in the past decades, an emerging research topic is the promising exploitation of the combination of structured knowledge bases (KBs) with neural networks (NNs) in view of seeking to overcome the limitations of each approach when taken separately. This has led to the research on hybrid systems that include retrieval-augmented neural models, Neuro-Symbolic systems and a range of other approaches for combining NNs and KBs.

Commonsense reasoning provides an interesting challenge in which the capabilities of Neuro-Symbolic methods can be matched up against Neuro-only methods. The reason for this is twofold. On the one hand, the universe covered by commonsense knowledge is so vast that it is currently unfeasible to provide a training dataset upon which to build a NN model that is able to cope with all that vastness and arrive at the relevant generalizations. On the other hand, though KBs are also far from complete, they capture prominent commonsense generalizations that can be usefully built upon by hybrid systems since such generalizations, by virtue of being commonsense, are often not explicitly stated in texts, making it hard for deep learning methods to learn them from raw textual data alone.

A recent paper [1] provides an exploratory study on commonsense reasoning tasks, with one of the findings suggesting that KBs may have little impact on the downstream task performance. A promising Neuro-Symbolic method named COMET [2, 3] was experimented with, which injects knowledge into the parameters of a network through a text generation task.

Against this background, we devise a broader experimentation setting aimed at empirically assessing, for commonsense reasoning, how promising and effective can be Neuro-Symbolic systems compared to Neuro-only systems, by covering different tasks and model types. The experimental space was defined by following these steps:

- Selecting four prominent NLP tasks in common sense reasoning. Most commonsense reasoning

KINN 2021: Workshop on Knowledge Injection in Neural Networks – November, 2021
rmbranco@fc.ul.pt (R. Branco); ambranco@fc.ul.pt (A. Branco); jrsilva@fc.ul.pt (J. Silva); jarodrigues@fc.ul.pt (J. Rodrigues)

tasks fall under two categories: multiple choice question answering (Q&A) or machine reading comprehension. In this exercise, we concentrated on Q&A commonsense reasoning tasks, covering different topics and different reasoning types, setting a demanding environment to probe the knowledge and ability of different language models.
- Selecting most prominent examples of non-hybrid, Neuro-only models from the three different families of Transformers, namely encoders-only, encoders-decoders and decoders-only; the models are, respectively, RoBERTa [4], T5 [5] and GPT-2 [6].
- Running these three models over the four selected tasks, finding the best performing one.
- Selecting a most promising hybrid system, namely COMET, which leverages a BART [7] model and enriches it with knowledge coming from a given KB. The simplicity of the injecting task and the interesting results obtained on generating commonsense knowledge makes COMET a promising method. We also include a fifth model, BART-Large, in the experimentation, serving as a baseline for COMET(BART).
- Finally, running this hybrid system over the same four tasks and comparing its performance against the best Neuro-only found above.

In this paper, we report on this experimental exercise and its results, and discuss the empirical evidence gathered in view of the research question motivated above: in what concerns commonsense reasoning, is there empirical evidence that hybrid models stand to the promise of having the potential of surpassing the performance of Neuro-only systems?

The results of our experimentation reported in the present paper do not uncover a clear winner, as in the performance scores obtained with the two approaches, no one stands out as having a clear advantage over the other, which comes in line with the findings from [1].[1]

2. Related Work

The Transformer model [8] has become the all-embracing and flexible approach to a wide range of NLP tasks. Extensive research has been devoted to refine its architecture and training methodology. It has become commonplace to section the training regime into two stages [9, 10, 11, 12], namely (i) **pre-training**, where a language model is trained on a large corpus of raw text, endowing it with some capacity to suitably process natural language (or at least a good amount of linguistic phenomena); and (ii) **fine-tuning**, where the model is refined with respect to a specific language processing task.

Ensuing works would improve the methodology, introducing different pre-training tasks that aid downstream performance [4, 5, 7]. These efforts have allowed an accelerating increase in the state of the art on benchmarks such as GLUE [13] and SuperGLUE [14]. Large pre-trained language models were shown to be competitive with systems that access external knowledge, in open-book Q&A challenges which require mostly trivia knowledge, by accessing their internal "memory" learned during pre-training [15]. These results raise the question of whether pre-training alone can endow language models with enough commonsense knowledge to tackle commonsense reasoning tasks.

Motivated by this question, a study [16] was conducted to assess the intrinsic knowledge and reasoning capabilities of different pre-trained Transformer-based language models, without fine-tuning. The models were applied to different commonsense reasoning tasks, by framing them as sentence scoring tasks. This can be achieved in the following manner: for a given example (e.g. a question and several possible answers), the perplexity of the language model when presented a particular question-answer pair is measured, and the pair with the lowest perplexity (thus being the one that makes the most sense) is considered to be the response of the model. The method requires no fine-tuning, allowing "probing" into the knowledge contained in the model. It was found that the models consistently performed better than random, albeit close to it in some tasks, and that a noticeable gap remains between the models and the human baselines.

A promising approach to aid the learning process of commonsense reasoning tasks are Neuro-Symbolic models. This approach attempts to combines knowledge stored in KBs with neural networks, providing a richer prior knowledge.

Early attempts at enriching neural networks with knowledge coming from KBs made use of Graph Embedding techniques, representing concepts as vectors which would be used to initialize the embedding layer. Techniques such as Node2Vec [17], TransE [18] and WordNet Embeddings [19, 20] can leverage the nature of graph structures to produce embeddings for the nodes through the strength of the relations between them. Lexical knowledge coming from ontologies such as WordNet can be used as a way to enrich neural networks in a Neuro-Symbolic system [21].

More recently, with the advent of the Transformer, methods to blend KBs within its parameters leveraged the self-attention system. Embeddings from external KBs are combined with the internal states from the Transformer

[1] Code is publicly available at: https://github.com/nlx-group/Commonsense-Reasoning-Neuro-only-vs-Neuro-Symbolic-Methods.

block, following some fusion function [22, 23, 24]. This usually requires the addition of a module that selects relevant entities, fetches their embeddings and includes them in a self-attention operation with the sequence hidden states.

These approaches are limited in three ways: (i) additional parameters required to be learned, making the search space for optimum performance rather expensive; (ii) given a trained model, accommodating additional entities, unseen during previous training, requires additional training, which may prove costly; and (iii) the embedding matrix of KB entities can quickly become a problem, especially as some popular and extensive KBs can encode millions of entities, being prohibitively expensive to hold in memory.

In an interesting new wave of research on commonsense reasoning inspired by GOFAI,[2], more complex KBs and methods are being developed to tackle the problem. Open resources for commonsense reasoning, whether manually built or automatically retrieved, tend to encode taxonomic knowledge, which is a subset of the types of commonsense knowledge. In ATOMIC [3, 25], we find an approach that seeks to improve reasoning by going beyond taxonomic knowledge to encoding causal and inferential knowledge. This is important for an agent/model to reason about what might be the causes for a certain event to happen, and given that it did, what we can infer from it. In its most recent update, ATOMIC encodes knowledge of social-interactions, physical-entity relations, and event-centered relations. ATOMIC is then used as a resource to build a dataset to fine-tune different generation models on a tail generation task, through a method named COMET (**Com**mons**e**nse **T**ransformers) [2, 3]. Tail generation task is designed to enable models to learn the knowledge contained in a KB. The task consists in presenting the model with a concept A and a relation, and the model must generate a concept B that has such relationship with concept A.

The reasoning behind the generation task lies in the fact that the universe of commonsense knowledge is so vast that we cannot hope to build a resource that is complete. Thus, we need our models to be able to generalize and generate new knowledge. The results show that without the use of the ATOMIC KB, pre-trained models fail to generate new knowledge, hinting at the possibility that the commonsense knowledge stored in them is very limited. When fine-tuned on the tail generation task, the COMET models are able to learn to generate new knowledge, even for entities that were not previously seen during training.

[2]Good Old-Fashioned AI

3. Tasks

We adopt four tasks related to commonsense reasoning, covering different demands and domains in terms of reasoning, framing a challenging environment to evaluate the capacity of models addressing them, and the effectiveness of the Neuro-Symbolic COMET method. The four tasks are (i) Argument Reasoning Comprehension, (ii) AI2 Reasoning Challenge, (iii) Physical Interaction Question Answering and (iv) CommonsenseQA.

3.1. Argument Reasoning Comprehension

The Argument Reasoning Comprehension Task (ARCT) [26] tests the argument reasoning ability of a model, requiring not only language and logic skills but also commonsense knowledge.

The underlying structure of an argument, whose uncovering dates back to Aristotle and his study of argumentation, is defined as a series of premises (reasons) that support a given claim (conclusion). In another model of argumentation, an additional fundamental part, named warrant, is included in the structure [27]. The warrant establishes the connection between the premises and claims, such that the latter must logically follow from the former (sequitur). Warrants are often implicit, under the assumption that they are shared knowledge between the addresser and addressee [28]. This makes identification of warrants an exercise that requires commonsense.

This task is defined as follows: given a reason and a claim, choose the appropriate warrant from two possible choices. One of the warrants is a distraction, not supporting the sequitur from reason to claim.

The dataset was constructed with data from the *Room for Debate*, a section in the New York Times,[3] where knowledgeable contributors participate in debates regarding contemporary issues. The authors selected 188 debates of controversial issues and used crowdworkers (referred as turkers) from Amazon Mechanical Turk to perform an 8-step pipeline to obtain the dataset instances, resulting in 1970 instances.

Problems of spurious correlations has been detected in the original dataset, so we will use a cleaned version of it [29].

3.2. AI2 Reasoning Challenge

The AI2 Reasoning Challenge (ARC) [30] is a multi-choice question answering task on the topic of natural sciences. The dataset is comprised of questions from 3rd to 9th-grade exams in the U.S. and other parts of the world. It is composed of two sets of questions: the easy and the

[3]https://www.nytimes.com/roomfordebate

challenge sets. The challenge set contains questions that cannot be trivially solved with token co-occurrence, as opposed to those in the easy set that can. Our experimentation will be carried out using the challenge set.

ARC demands models to possess knowledge in different dimensions: definitions, facts and properties, structure, processes and causal, teleology/purpose, algebraic, and many more; and different reasoning types: question logic, linguistic matching, multi-hop, comparison, algebraic, etc. The diversity in knowledge and reasoning types required to learn ARC makes it a highly challenging task.

3.3. Physical Interaction Question Answering

The Physical Interaction Question Answering task (PIQA) [31] tests the capabilities of models to answer commonsense questions regarding the physical world. Models are presented with a goal, mostly an everyday situation that a human might want to accomplish, and two possible solutions to attain the goal. Models will need to learn, from raw text only, physical commonsense knowledge.

For humans, acquiring physical commonsense knowledge is part of the human experience. We can interact with the world, manipulate objects and figure out how we might use them to solve problem, in a process called grounding [32]. Unlike humans, models as of yet cannot interact with the world to learn these properties, which makes it a real challenge for them to acquire physical knowledge from raw text only.

Large scale language models struggle with this task, with the state of the art achieving 83.5%[4] accuracy, compared to the human 95% score.

3.4. CommonsenseQA

CommonsenseQA (CSQA) [33] is a multi-choice question answering dataset that requires commonsense knowledge in different formats, akin to ARC. It encompasses many different knowledge types: spatial, cause and effect, has parts, is member of, purpose, social, activity, definition and preconditions.

It was built by resorting to ConceptNet [34], extracting subgraphs of concepts which are used to build questions, through crowdsourcing with Amazon Mechanical Turk.

4. Methodology

Designing a broad evaluation setting enables a richer comparison between the Neuro-only pre-trained language models and the Neuro-Symbolic model. For comparison, we pick a representative from each Transformer family of models:

Encoder-only We adopted RoBERTa [4] as an encoder-only exemplar. RoBERTa [4] is a derivative of BERT [12], conceptualized from a study on the optimization of BERT models.

Decoder-only The GPT series of models have gained notoriety in NLP, and we select the most recent *computationally affordable* version, GPT-2 [6]. GPT-2 is a left-to-right language model, comprised of stacked Transformer decoders. It excels in text generation and boasts considerable capabilities in Natural Language Understanding (NLU) tasks.

Encoder-Decoder For this family of architectures, we resorted to T5 [5]. T5 is conceptualized as a text-to-text framework, meaning that both the input and output are entirely textual, regardless of the underlying tasks. This affords T5 with immense flexibility.

To inject finer priors into a language model, in a Neuro-Symbolic approach, we follow the COMET [2] method, which leverages a generative task to enrich the model with a commonsense knowledge base. In the current paper, we use COMET(BART), which is a BART-Large [7] model trained with a tail generation task on ATOMIC2020 [3], the most recent version of the ATOMIC knowledge base. In order to better assess the influence of the knowledge base, we also include a standard BART-Large in our evaluation.

A different fine-tuning technique is used for RoBERTa, GPT-2, COMET(BART) and BART-Large, which was shown as promising for Q&A based commonsense reasoning tasks. The fine-tuning process frames problems into sequence ranking problems [35]. In this framing, given that the tasks are multiple-choice, the elements of input pairs of questions and candidate answers, (q_i, a_i), are separately given to the network, which produces a value, named the relevancy score. The pair with the maximum relevancy score is the answer given by the network.

Implementations for the experimentation are based on Huggingface [36], along with their pre-trained model weights.

A sequential hyper-parameter search is employed for the learning rate and batch size. First, the learning rate is determined through the selection of the model with the best accuracy on the development set, after fine-tuning for 10 epochs, with learning rate values picked from the set {1e-3, 1e-4, 1e-5, 2e-3, 2e-4, 2e-5, 3e-3, 3e-4, 3e-5}. An appropriate batch size is subsequently searched for, using the previously determined learning rate. The same

[4] Accessed on 2021/07/31: https://yonatanbisk.com/piqa/

Table 1
Accuracy (with standard deviation) on the selected tasks. Best result for each task in bold. Human benchmarks for CSQA obtained from their public leaderboard,[5] for ARCT from [26], and for PIQA from [31].

	ARCT	ARC	PIQA	CSQA	Params
Random	0.5	0.25	0.5	0.2	-
HUMAN	0.909	N/A	0.949	0.889	-
RoBERTa-Large	**0.815 ± 0.011**	0.411 ± 0.022	**0.789 ± 0.006**	0.733 ± 0.006	355M
GPT2-Medium	0.540 ± 0.071	0.318 ± 0.009	0.706 ± 0.005	0.551 ± 0.012	345M
T5-Large	0.743 ± 0.006	**0.440 ± 0.008**	0.772 ± 0.005	0.713 ± 0.007	770M
BART-Large	0.655 ± 0.154	0.382 ± 0.027	0.777 ± 0.005	**0.738 ± 0.005**	406M
COMET(BART)	0.790 ± 0.005	0.412 ± 0.011	0.783 ± 0.008	0.718 ± 0.008	406M

strategy is followed, searching for a batch size from the set {4, 8, 16, 32}. The hyper-parameters found for each model are described in Appendix A.

Models are trained for up to 30 epochs, and the checkpoint yielding the best accuracy on the development set is selected for testing purposes. Due to known instability across runs in pre-trained language models [12, 37], we report the mean of five runs, each with different random seeds, which are described in Appendix A.

PIQA and CSQA, two of our proposed tasks, are active competitions, and as such their test sets are kept private so as to prevent cheating. Thus, for these tasks, we report results on the development set. In order to have a training set and a development set available, we split the training set into two using stratified splitting to preserve the original distribution of classes, keeping 90% of the data as training data and setting aside 10% of the data as the development set.

All experiments were conducted on a single NVIDIA Titan RTX with 24Gb VRAM.

5. Results and Discussion

Performance (in accuracy) for the commonsense reasoning tasks is shown in Table 5.

A gap between human performance and the performance of the models is noticeable, albeit some tasks standout as more challenging for the models than others. For ARCT, an encouraging gap of 0.094 accuracy separates the human upper bound from RoBERTa, which is the best performing model for the task. CSQA and PIQA have a more significant margin to the human upper bound, with a gap of 0.156 and 0.151, respectively. Despite the advances provided by Transformer and the "pre-train then fine-tune" methodology, which have pushed the state of the art further and the boundaries of sizes of model and training data, models are still a long way away from human aptitude.

A juxtaposition appears concerning ARC and CSQA. Despite both being multiple-choice problems with (up to) five possible answers, scores on ARC are half of those on CSQA. The commonsense knowledge types for CSQA is more broad, covering a large array of domains, being a more general task. ARC, on the other hand, focuses on more complex knowledge about the physical world, as it was conceived from science exams, and as such features knowledge about physics and chemistry. This hints at the possibility that ARC is a harder task to solve.

Regarding our primary research question, which is to find empirical clues as to whether COMET effectively boosts performance on commonsense reasoning tasks, the answer is that it is slightly better than its baseline. COMET(BART), outperforms its baseline BART-Large on all but the CSQA task. However, on the ARCT task, the standard deviation of BART-Large is so large due to instability that, despite the differences in the mean, it is unclear whether COMET(BART) provides any real advantage in this task.[6] As such, given that BART-Large excels on CSQA, and on ARCT it could be that BART-Large can perform at the level of COMET(BART), if it were not for the instability, COMET does not provide a sizeable difference over the baseline, but it does improve over it.

When comparing COMET(BART) with the different Transformer families, COMET(BART) consistently holds the 2nd and 3rd best scores, reaching close to the best score on PIQA. It was competitive with T5 (Encoder-Decoder) and GPT-2 (Decoder-only), but scored below RoBERTa (Encoder-only) in all tasks but ARC. RoBERTa, while not having had the knowledge injection for finer priors, still emerges as the most competent reasoner, having the best score in ARCT and PIQA, and is a close second best in CSQA. The worst reasoner was GPT-2, obtaining poor results in all tasks.

Despite the ability of the Neuro-Symbolic COMET(BART) to be competitive with most Neuro-only

[5] https://www.tau-nlp.org/csqa-leaderboard

[6] A t-test between COMET(BART) and BART-Large yielded $p > 0.05$ for the ARCT, ARC and PIQA results.

methods, no clear advantage can be given to any of the two approaches overall, for the proposed tasks. This finding confirms previous work [1], which has reached the same conclusion for other tasks. One particular Neuro-only method stood out, RoBERTa, which consistently outperformed other Neuro-only methods and COMET(BART). RoBERTa was pre-trained on the same corpus as BART, and as such, COMET(BART) was exposed to the same data and more, due to its refinement on the ATOMIC knowledge base, meaning differences in the training data should not be a factor in the differences between RoBERTa's and COMET(BART)'s performance. We conjecture that two factors, alone or in combination, could account for the observed differences:

- The generative nature of COMET's pre-training task, while being ideal for **generating** new commonsense knowledge, proves to be detrimental to classification tasks, in this instance, for commonsense reasoning.
- RoBERTa's architecture, featuring only the Encoder, could be better adapted to perform commonsense reasoning. Perhaps, a combination of COMET with RoBERTa could yield better results.[7]

Overall, the models behaved like capable reasoners, performing well above the random baseline, despite the gap to human capabilities. One can pose the question of whether this performance is consistent across different types of reasoning in the different tasks. We have explored the question of consistency in previous work [38]. We have found evidence that models are inconsistent because they seem to be solving the tasks not with reasoning but with shortcuts present in the data. In one instance, we removed the question portion of the input from the tasks, and the models were still able to learn to select the correct answers to questions they were not presented with. Adversarial attacks also demonstrate that minimal superficial changes to the input have a significant impact on their performance. COMET(BART) was just as susceptible to shortcuts as other models. However, one can expect that with further research into neuro-symbolism, the enrichment procedure should endow models with finer priors such that they should become more resistant to this behavior.

6. Conclusion

Neuro-Symbolic methods are a promising path to integrate the rich knowledge represented in structured knowledge bases with deep learning models, enhancing their learning capabilities.

[7] Assuming an adaptation of COMET's tail generation task to a classification task is possible.

In this work, we establish a broad and challenging evaluation setting to gauge the efficiency of a Neuro-Symbolic method, COMET, at learning and applying the knowledge learned from the ATOMIC2020 knowledge base, and how it compares to Neuro-only methods. Five models (RoBERTa, GPT-2, T5, BART and COMET(BART)), covering the three families of Transformer models (Encoder, Decoder and Encoder-Decoder), are tasked with four challenging commonsense reasoning tasks (ARCT, ARC, PIQA and CSQA).

The results show no clear advantage between Neuro-only and Neuro-Symbolic methods. COMET(BART) is marginally better than its BART-Large baseline, and is competitive with most Neuro-only methods. RoBERTa emerges as the superior reasoner, despite not having been afforded with finer priors like COMET(BART).

These results call for future research on two different topics: (i) The application of the COMET method to other models in the vast Transformer family. Despite COMET being inherently a generative method, its adaptation to models like RoBERTa, which has shown great promise in our results, could prove to be beneficial; and (ii) An intensive systematic review spanning different Neuro-Symbolic methods, each with their particular techniques, and their application to commonsense reasoning. This review would help uncover what type of techniques more effectively transfer knowledge to models, and aid their learning.

Acknowledgments

The research reported here was supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language https://portulanclarin.net, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

References

[1] N. Lourie, R. Le Bras, C. Bhagavatula, Y. Choi, Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark, AAAI (2021).

[2] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, Comet: Commonsense transformers for automatic knowledge graph construction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4762–4779.

[3] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs, in: AAAI'21, 2021.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[9] A. M. Dai, Q. V. Le, Semi-supervised sequence learning, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 28, Curran Associates, Inc., 2015.

[10] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339.

[11] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018.

[12] J. Devlin, M.-W. Chang, K. Lee, K. 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[13] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, pp. 353–355.

[14] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, SuperGLUE: A stickier benchmark for general-purpose language understanding systems, arXiv preprint 1905.00537 (2019).

[15] A. Roberts, C. Raffel, N. Shazeer, 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), Association for Computational Linguistics, Online, 2020, pp. 5418–5426.

[16] X. Zhou, Y. Zhang, L. Cui, D. Huang, Evaluating commonsense in pre-trained language models., in: AAAI, 2020, pp. 9733–9740.

[17] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.

[18] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Neural Information Processing Systems (NIPS), 2013, pp. 1–9.

[19] R. Branco, J. Rodrigues, C. Saedi, A. Branco, Assessing wordnets with WordNet embeddings, in: Proceedings of the 10th Global Wordnet Conference, Global Wordnet Association, Wroclaw, Poland, 2019, pp. 253–259.

[20] C. Saedi, A. Branco, J. Rodrigues, J. Silva, Wordnet embeddings, in: Proceedings of the third workshop on representation learning for NLP, 2018, pp. 122–131.

[21] M. Salawa, A. Branco, R. Branco, J. António Rodrigues, C. Saedi, Whom to learn from? graph- vs. text-based word embeddings, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 1041–1051.

[22] M. E. Peters, M. Neumann, R. L. Logan, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, in: EMNLP, 2019.

[23] A. Yang, Q. Wang, J. Liu, K. Liu, Y. Lyu, H. Wu, Q. She, S. Li, Enhancing pre-trained language representations with rich knowledge for machine reading comprehension, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2346–2357.

[24] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, Ernie: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451.

[25] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, in: Proceedings of the AAAI Con-

ference on Artificial Intelligence, volume 33, 2019, pp. 3027–3035.
[26] I. Habernal, H. Wachsmuth, I. Gurevych, B. Stein, The argument reasoning comprehension task: Identification and reconstruction of implicit warrants, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1930–1940.
[27] S. E. Toulmin, The Uses of Argument, Cambridge University Press, 1958.
[28] J. B. Freeman, Argument Structure:: Representation and Theory, volume 18, Springer Science & Business Media, 2011.
[29] T. Niven, H.-Y. Kao, Probing neural network comprehension of natural language arguments, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4658–4664.
[30] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, arXiv preprint arXiv:1803.05457 (2018).
[31] Y. Bisk, R. Zellers, R. LeBras, J. Gao, Y. Choi, Piqa: Reasoning about physical commonsense in natural language., in: AAAI, 2020, pp. 7432–7439.
[32] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, J. Turian, Experience grounds language, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8718–8735.
[33] A. Talmor, J. Herzig, N. Lourie, J. Berant, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4149–4158.
[34] H. Liu, P. Singh, Conceptnet—a practical commonsense reasoning tool-kit, BT technology journal 22 (2004) 211–226.
[35] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4487–4496.
[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45.
[37] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines, in: International Conference on Learning Representations, 2021.
[38] R. Branco, A. Branco, J. Silva, J. Rodrigues, Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2021.

A. Training Hyper-Parameters

The random seeds used for evaluation purposes were the following: 42, 1128, 1143, 1385 and 1415.

The table below describes the hyper-parameters (Batch Size, Learning Rate, Epochs) used in the experiments.

Task	Model	Hyper-parameters
ARCT	RoBERTa-Large	16, 1e-5, 25
	GPT2-Medium	8, 2e-3, 18
	T5	8, 2e-5, 17
	BART-Large	16, 2e-4, 12
	COMET(BART)	8, 1e-4, 25
ARC	RoBERTa-Large	8, 1e-4, 16
	GPT2-Medium	4, 1e-3, 26
	T5	8, 2e-5, 12
	BART-Large	8, 1e-4, 27
	COMET(BART)	8, 3e-5, 22
PIQA	RoBERTa-Large	16, 3e-3, 28
	GPT2-Medium	8, 1e-3, 22
	T5	8, 1e-5, 9
	BART-Large	4, 1e-3, 19
	COMET(BART)	32, 3e-4, 16
CSQA	RoBERTa-Large	8, 3e-4, 13
	GPT2-Medium	8, 1e-3, 14
	T5	8, 2e-5, 5
	BART-Large	8, 3e-4, 18
	COMET(BART)	8, 1e-4, 14