Effective Cross-Task Transfer Learning for Explainable Natural Language Inference with *T5*

Irina Bigoulaeva[1*], Rachneet Sachdeva[1*], Harish Tayyar Madabushi[2*], Aline Villavicencio[3] and Iryna Gurevych[1]

[1] Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt
[2] Department of Computer Science, The University of Bath
[3] Department of Computer Science, The University of Sheffield

www.ukp.tu-darmstadt.de

htm43@bath.ac.uk, a.villavicencio@sheffield.ac.uk

Abstract

We compare sequential fine-tuning with a model for multi-task learning in the context where we are interested in boosting performance on two tasks, one of which depends on the other. We test these models on the *FigLang2022* shared task which requires participants to predict language inference labels on figurative language along with corresponding textual explanations of the inference predictions. Our results show that while sequential multi-task learning can be tuned to be good at the first of two target tasks, it performs less well on the second and additionally struggles with overfitting. Our findings show that simple sequential fine-tuning of text-to-text models is an extraordinarily powerful method for cross-task knowledge transfer while simultaneously predicting multiple interdependent targets. So much so, that our best model achieved the (tied) *highest score* on the task[1].

1 Introduction and Motivation

The transfer of information between *supervised learning objectives* can be achieved in Pre-trained Language Models (PLMs) using either multi-task learning (MTL) (Caruana, 1997) or sequential fine-tuning (SFT) (Phang et al., 2018). MTL involves simultaneously training a model on multiple learning objectives using a weighted sum of their loss, while SFT involves sequentially training on a set of related tasks. Recent work has extended the SFT approach by converting all NLP problems into text-to-text (i.e., sequence-to-sequence where both input and output sequences are natural text) problems (Raffel et al., 2019). The resultant model – *T5* – has achieved state-of-the-art results on a variety of tasks such as question answering, sentiment analysis, and, most relevant to this work, Natural Language Inference (NLI).

In this work, we focus our efforts on the transfer of information from multiple related tasks for improved performance on a different set of tasks. In addition, we compare the effectiveness of SFT with that of MTL in a context where one of the target tasks is dependent on the other. Given the dependence of one of the target tasks on the other, we implement an end-to-end multi-task learning model to perform each of the tasks sequentially: an architecture referred to as a *hierarchical feature pipeline* based MTL architecture (*HiFeatMTL*, for short) (Chen et al., 2021). While HiFeatMTL has been previously used in different contexts (see Section 3), it has, to the best of our knowledge, *not* been used with, or compared to, text-to-text models. This is of particular importance as such models are known to enable transfer learning (Raffel et al., 2019) and it is crucial to determine if traditional MTL methods can boost cross-task knowledge transfer in such models.

Specifically we participate in the FigLang2022 Shared Task[2], which extends NLI to include a figurative-language hypothesis and additionally requires participants to output a textual explanation (also see Section 2). FigLang2022 is ideally suited for the exploration of knowledge transfer, as PLMs have been shown to struggle with figurative language and so any gains achieved are a result of knowledge transfer. For example, Liu et al. (2022) show that in the zero- and few-shot settings, PLMs perform significantly worse than humans. This is especially the case with idioms (Yu and Ettinger, 2020; Tayyar Madabushi et al., 2021), on which T5 does particularly poorly (see Section 4). Additionally, FigLang2022's emphasis on explanations of the predicted labels provides us with the oppor-

[*] Equal Contribution

[1] To ensure reproducibility and to enable other researchers to build upon our work, we make our code and models freely available at https://github.com/Rachneet/cross-task-figurative-explanations

[2] https://figlang2022sharedtask.github.io/

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 54 - 60

tunity to test cross-task knowledge transfer in a setting where one target task depends on the other (HiFeatMTL) – this is especially so given the evaluation methods used (detailed in Section 2).

We evaluate the effectiveness of boosting performance on the target tasks through the transfer of information from two related tasks: a) eSNLI, which is a dataset consisting of explanations associated with NLI labels, and b) IMPLI, which is an NLI dataset (without explanations) that contains figurative language. More concretely, we set out to answer the following research questions:

1. Can distinct task-specific knowledge be transferred from separate tasks so as to improve performance on a target task? Concretely, can we transfer explanations of literal language from eSNLI and figurative NLI without explanations from IMPLI?
2. Which of the two knowledge transfer techniques (SFT or HiFeatMTL) is more effective in the text-to-text context?

2 The FigLang2022 Shared Task

FigLang2022 is a variation of the NLI task which requires the generation of a textual explanation for the NLI prediction. Additionally, the hypothesis is a sentence that employs one of four kinds of figurative expressions: *sarcasm, simile, idiom,* or *metaphor*. Additionally, a hypothesis can be a *creative paraphrase*, which rewords the premise using more expressive, literal terminology. Table 1 shows examples from the task dataset.

Entailment	
Premise	I respectfully disagree.
Hypothesis	I beg to differ. *(Idiom)*
Explanation	To beg to differ is to disagree with someone, and in this sentence the speaker is respectfully disagreeing.
Contradiction	
Premise	She was calm.
Hypothesis	She was like a kitten in a den of coyotes. *(Simile)*
Explanation	A kitten in a den of coyotes would be scared and not calm.

Table 1: An entailment and contradiction pair from the FigLang2022 dataset.

FigLang2022 takes into consideration the quality of the generated explanation when assessing the model's performance by use of an *explanation score*, which is the average between BERTScore and BLEURT and ranges between 0 and 100. The task leaderboard is based on NLI label accuracy at an explanation score threshold of 60, although the NLI label accuracy is reported at three thresholds of the explanation score (i.e. 0, 50, and 60) so as to provide a glimpse of how the model's NLI and explanation abilities are influenced by each other.

3 Related Work

NLI is considered central to the task of Natural Language Understanding, and there has been significant focus on the development of models that can perform well on the task (Wang et al., 2018). This task of language inference has been independently extended to incorporate explanations (Camburu et al., 2018) and figurative language (Stowe et al., 2022) (both detailed below). Chakrabarty et al. (2022) introduced *FLUTE*, the Figurative Language Understanding and Textual Explanations dataset which brought together these two aspects.

Previous shared tasks involving figurative language focused on the identification or representation of figurative knowledge: For example, FigLang2020 (Klebanov et al., 2020) and Task 6 of SemEval 2022 (Abu Farha et al., 2022) involved sarcasm detection, and Task 2 of SemEval 2022 (Tayyar Madabushi et al., 2022) involved the identification and representation of idioms.

The generation of textual explanations necessitates the use of generative models such as BART (Lewis et al., 2020) or T5 (Raffel et al., 2019). Narang et al. (2020) introduce WT5, a sequence-to-sequence model that outputs natural-text explanations alongside its predictions and Erliksson et al. (2021) found T5 to consistently outperform BART in explanation generation.

Of specific relevance to our work are the IMPLI (Stowe et al., 2022) and eSNLI (Camburu et al., 2018) datasets. IMPLI links a figurative sentence, specifically idiomatic or metaphoric, to a literal counterpart, with the NLI relation being either entailment or non-entailment. Stowe et al. (2022) show that idioms are difficult for models to handle, particularly in non-entailment relations. The eSNLI dataset (Camburu et al., 2018) is an explanation dataset for general NLI. It extends the Stanford Natural Language Inference dataset (Bowman et al., 2015) with human-generated text explanations.

Hierarchical feature pipeline based MTL architectures (*HiFeatMTL*) use the outputs of one task as a feature in the next and are distinct from hierarchical *signal* pipeline architectures wherein the

outputs are used indirectly (e.g., their probabilities) (Chen et al., 2021). HiFeatMTL has previously been used variously (Fei et al., 2019; Gong et al., 2019; Song et al., 2020), including, for example, to provide PoS and other syntactic information to relatedness prediction, the output of which is, in addition to the syntactic features, passed to an entailment task (Hashimoto et al., 2017) (see also the survey by Chen et al. (2021)). To the best of our knowledge, this is the first work to use HiFeatMTL with, and to compare against, text-to-text models and their ability to transfer knowledge across tasks.

4 Methods

We set out to answer the research questions in Section 1 by evaluating the effectiveness of SFT and HiFeatMTL on the transfer of task-specific knowledge from separate tasks, namely, explanations from eSNLI and figurative language from IMPLI. We use T5 for all our experiments as it has been shown to be effective in explanation generation (Erliksson et al., 2021). We run all our hyperparameter optimisation and model variations using T5-base (evaluated on a development split consisting of 10% of the training data) before then transferring over the best performing settings to T5 large (trained on all of the training data) which is used to make predictions on the test set. While we find this method adequate in finding a good set of hyperparameters, the best setting for a smaller model need not necessarily be a good setting for larger models, especially given that some capabilities emerge only in larger models (Wei et al., 2022).

4.1 Exploratory Experiments

The first phase of our experiments was dedicated to using our development split to determining the best hyperparameters for T5, specifically the learning rate, and the number of beams, the two parameters that we found T5 to be extremely sensitive to. We do not experiment with prompt optimisation, but rather our prompts are based on what T5 was trained on (See listing 1).

```
Source_text:
    figurative hypothesis: <hypothesis> premise:
        <premise>
target_text:
    <label> explanation: <explanation>
```

Listing 1: Our default prompt used for T5.

An additional consideration of this initial phase was whether it was more effective to independently perform the task of NLI before subsequently generating explanations. However, we find that incorporating the gold inference labels does not improve the quality of explanations generated.

Knowledge Transfer To determine those forms of figurative language that T5 finds challenging and how effective knowledge transfer is, we test T5 fine-tuned just on FigLang2022, and sequentially on IMPLI followed by FigLang2022. The results of these experiments are presented in Table 2, which correspond to the observations made by Stowe et al. (2022) that idioms are particularly challenging for NLI models. Crucially, we find that the performance of the model *does* improve when first trained on IMPLI, thus establishing that knowledge transfer is possible in T5 through SFT.

Type	FigLang	IMPLI → FigLang
Metaphor	81.97	**83.61 (+ 2.0%)**
Simile	**65.38**	**66.92 (+ 1.5%)**
Idioms	**72.50**	**78.13 (+ 6.0%)**
Creative Paraphrase	98.36	98.36
Sarcasm	100	99.54 (- 0.5%)

Table 2: T5 performance (acc) on the various labels of FigLang2022, before and after training on IMPLI.

Importantly, we found that training for more epochs on the IMPLI dataset led to improved inference label accuracy but led to poorer explanations, which suggests knowledge transfer as oppsed to, for example, the advantage of additional training data. Since we were more interested in transferring figurative information from IMPLI, we optimise on Acc@0 (label accuracy) when training on IMPLI and Acc@60 (the evaluation metric relevant to the task) when training on the final FigLang dataset.

4.2 Experimental Setup

Training Regime In establishing the most effective method of knowledge transfer, we compare SFT with HiFeatMTL trained on: a) FigLang, b) eSNLI → FigLang, c) IMPLI → FigLang, d) eSNLI → IMPLI → FigLang, and e) IMPLI → eSNLI → FigLang. The training sets of both eSNLI and IMPLI are truncated to the same length as that of FigLang to ensure that the model does not over-fit on those other tasks.

4.3 Sequential Fine-Tuning

In SFT, we fine-tune the model on each of the relevant datasets in sequence. When training on the IMPLI dataset, which does not have associated explanations, we use the same prompt (Listing 1) but with no associated explanation. The number of

training epochs is established based on the change in loss on the development set and was found to be 3 for IMPLI and 10 for the other two datasets.

4.4 Multi-Task Learning

We experiment with a *hierarchical feature pipeline* for multi-task learning as the output inference label is likely to be important in generating the explanation. This involved creating an end-to-end model wherein, during the forward pass, T5 is used to predict the inference labels based on the hypothesis and the premise. This label, in addition to the hypothesis and premise are then used as input to T5 to generate an explanation. During the backward pass, the overall loss of the model is calculated as the weighted sum of the loss associated with each of the two steps above. Importantly, the weights of the T5 model used in the two steps are shared. Figure 1 provides an illustration.

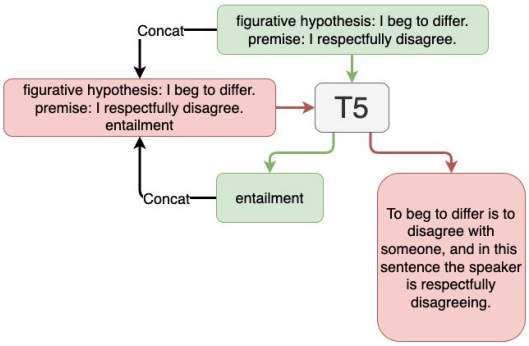

Figure 1: Our HiFeatMTL architecture. Note that we do not use GPT in our experiments, although it is possible to use GPT in place of T5.

As in the case of SFT, we fine-tune the model on each of the relevant datasets in sequence. When training on the FigLang dataset, we found it effective to train the model twice: fist with a higher weight to the loss associated with the inference (90%) and a second time with a higher weight to the loss associated with explanations (also 90%). Due to the summing of losses, we found that the model loss was not a good indicator of overfitting and instead determined the number of training epochs experimentally (10 for all datasets).

5 Results and Discussion

Table 3 shows the full shared task results from the CodaLab leaderboard[3] as of the competition's end date of 20 Aug, 2022. Our results (Team UKPChefs) are highlighted in bold.

Rank	Team Name	Acc@0	Acc@50	Acc@60
***1st**	**UKPChefs**	**0.925**	**0.869**	**0.633**
*1st	TeamCoolDoge	0.947	0.889	0.633
2nd	vund	0.936	0.865	0.607
3rd	hoho5702	0.911	0.854	0.548
4th	yklal95	0.847	0.779	0.517
5th	tuhinnlp	0.443	0.443	0.443
6th	peratham.bkk	0.590	0.203	0.033
Shared Task Baseline		*0.817*	*0.748*	*0.483*

Table 3: Shared task results from all teams (ours – UKPChefs – in bold). Asterisks represent tied results.

The results of our experiments using SFT and HiFeatMTL are presented in Table 4. The results on the development set and those on the test set are not directly comparable: not only do we use different models, we also train on all the complete training data before evaluating on the test set. The drop in performance of the HiFeatMTL model on the test set on Acc@60, which consistently outperforming SFT on Acc@0 across both the development and the test sets is surprising. This seems to indicate that HiFeatMTL, while an effective way of boosting performance on the earlier of multiple dependent objectives, seems to be less effective on subsequent tasks (in this case, explanation generation). Additionally, HiFeatMTL also seems prone to overfitting, as the FigLang test set introduced novel idioms and similes previously unseen in the training set, into the test set.

While the gain in accuracy when using the additional datasets could be due to the corresponding addition of training data, it should be noted that IMPLI does not have explanations and eSNLI contains no figurative language. As such, the improved scores indicate the transfer of figurative information from one task (IMPLI) and explanation generation capabilities from another (eSNLI).

As such, in addressing the research questions, our results indicate that: a) distinct task-specific knowledge (i.e. explanations or figurative language) can indeed be transferred from separate tasks so as to improve performance on a target task, and b) SFT seems to be a more effective way of transferring knowledge across tasks when we are concerned with the latter of a sequence of tasks (as in this case), while HiFeatMTL seems effective in boosting the performance of the first.

[3]Our CodaLab submissions appear under the name "rachneet": https://codalab.lisn.upsaclay.fr/competitions/5908

	Dataset 1	Dataset 2	Dataset 3	Acc@0		Acc@50		Acc@60	
				Dev	Test	Dev	Test	Dev	Test
SFT	FigLang	-	-	84.99	93.27	78.49	87.80	56.18	61.74
	eSNLI	FigLang	-	86.06	92.67	80.74	87.20	57.77	63.27
	IMPLI	FigLang	-	86.59	93.20	80.74	87.33	56.97	60.93
	eSNLI	IMPLI	FigLang	86.32	92.47	80.08	86.87	58.17	63.33
	IMPLI	eSNLI	FigLang	84.99	92.73	79.42	87.33	55.38	62.00
HiFeatMTL	FigLang	-	-	91.24	94.67	82.07	86.54	55.11	55.13
	eSNLI	FigLang	-	91.50	94.14	82.07	86.40	55.91	53.80
	IMPLI	FigLang	-	89.50	N/A	81.27	N/A	55.78	N/A
	eSNLI	IMPLI	FigLang	90.97	94.54	80.35	85.94	53.92	54.27
	IMPLI	eSNLI	FigLang	89.37	N/A	80.34	N/A	53.52	N/A
Shared Task Baseline				-	*81.70*	-	*74.80*	-	*48.30*

Table 4: Results of the SFT and HiFeatMTL models on the development and test splits of the FigLang2022 task. Experiments on the dev set were performed using T5-Base and those on the test set on T5-Large trained on the complete training set. Results marked N/A were not obtained due to the limits on the number of submissions.

6 Knowledge Transfer vs Bias

Recent works on NLI have shown that for some datasets, models are able to correctly predict the label using only the hypothesis, without considering the premise (Glockner et al., 2018; Gururangan et al., 2018; McCoy et al., 2019). This is caused by the model exploiting spurious correlations or patterns in the data, rather than acquiring task-relevant knowledge. As such, we wish to analyse if this is the case with our models: namely, whether our models employ figurative language knowledge from the hypothesis when predicting NLI labels.

We perform the following experiments using T5 large on our validation set: we train only the hypothesis, only on the premise, and compare these results with a model trained on both (the standard training regime). The results (Table 5) indicate that, while the model *can* achieve reasonable accuracy while relying solely on the hypothesis, the significant improvement in accuracy (on both Acc@0 and Acc@60) when considering both the hypothesis and the premise indicates that, to a certain extent, the model is using knowledge of figurative language to predict the NLI labels and corresponding explanations.

Setting	Acc@0	Acc@50	Acc@60
Regular	92.16	87.92	66.14
Hyp-Only	65.47	60.96	45.95
Prem-Only	56.31	47.81	33.74

Table 5: T5-large performance on the FigLang dataset with either the hypothesis or premise removed.

7 Conclusions and Future work

In this work we set out to establish the possibility of effectively transferring knowledge across tasks in the context where we are interested in boosting the performance of two dependent tasks. As such, we evaluate the effectiveness of SFT and HiFeatMTL for transferring distinct task-specific knowledge from different tasks and find that both of these methods are good at achieving this: SFT on the last task and HiFeatMTL on the first. We find that using SFT to transfer information across tasks is, in this instance, so effective that we are *ranked first* on the FigLang 2022 task.

In extending this work, we intend to test these methods on a variety of sequentially dependent tasks as well as incorporating the use of more efficient MTL methods including AdapterFusion (Pfeiffer et al., 2021) and AdapterDrop (Rücklé et al., 2021).

Acknowledgements

This work was made possible through a research visit hosted by the UKP Lab[4] and funded by the Alan Turing Institute[5] through their Post-Doctoral Enrichment Award granted to HTM while at the University of Sheffield. In addition, this work was also partly supported by the UK EPSRC grant EP/T02450X/1, the European Regional Development Fund (ERDF), the Hessian State Chancellery – Hessian Minister of Digital Strategy and Development (reference 20005482, TexPrax), the State of Hesse in Germany (project 71574093, CDR-CAT), the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

[4] https://www.informatik.tu-darmstadt.de/ukp/ukp_home/

[5] https://www.turing.ac.uk/

Limitations

This work only deals with English, and since English makes up a majority of the training data for PLMs, performance may drop across other languages. Additionally, we only address figurative language within the context of the NLI task, and thus do not make broader claims about our model's ability to handle figurative language, to generate explanations or generalise across other generative models. This also extends to the comparisons between models that we present.

Model Explanations This work is involved in the generation of explanations associated with language inference predictions. Importantly, there is no guarantee (and very unlikely) that the generated explanations are indeed faithful to the process of predicting inference labels (also see Jacovi and Goldberg (2020)).

Carbon Footprint All initial experiments are performed on smaller models and the best performing model architectures and parameters are transferred over to larger models to minimise the carbon footprint of our experiments. Despite this, the use of large language models does contribute to the climate crisis.

References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding and textual explanations.

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-task learning in natural language processing: An overview. *CoRR*, abs/2109.09138.

Karl Fredrik Erliksson, Anders Arpteg, Mihhail Matskin, and Amir H Payberah. 2021. Cross-domain transfer of generative explanations using text-to-text models. In *International Conference on Applications of Natural Language to Information Systems*, pages 76–89. Springer.

Hongliang Fei, Shulong Tan, and Ping Li. 2019. Hierarchical multi-task word embedding learning for synonym prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 834–842, New York, NY, USA. Association for Computing Machinery.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yu Gong, Xusheng Luo, Yu Zhu, Wenwu Ou, Zhao Li, Muhua Zhu, Kenny Q. Zhu, Lu Duan, and Xi Chen. 2019. Deep cascade multi-task learning for slot filling in online shopping assistant. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

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

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna

Feldman, and Debanjan Ghosh, editors. 2020. *Proceedings of the Second Workshop on Figurative Language Processing*. Association for Computational Linguistics, Online.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2020. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3875–3881. International Joint Conferences on Artificial Intelligence Organization. Main track.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.