On the Relationship between Sentence Analogy Identification and Sentence Structure Encoding in Large Language Models

Thilini Wijesiriwardene[1]*, Ruwan Wickramarachchi[1], Aishwarya Naresh Reganti[2]
Vinija Jain[3,4†], Aman Chadha[3,4†], Amit Sheth[1], Amitava Das[1]
[1]AI Institute, University of South Carolina, USA,
[2]Carnegie Mellon University, Pittsburgh, USA,
[3]Amazon GenAI, USA, [4]Stanford University, USA
thilini@sc.edu

Abstract

The ability of Large Language Models (LLMs) to encode syntactic and semantic structures of language is well examined in NLP. Additionally, analogy identification, in the form of word analogies are extensively studied in the last decade of language modeling literature. In this work we specifically look at how LLMs' abilities to capture sentence analogies (sentences that convey analogous meaning to each other) vary with LLMs' abilities to encode syntactic and semantic structures of sentences. Through our analysis, we find that LLMs' ability to identify sentence analogies is positively correlated with their ability to encode syntactic and semantic structures of sentences. Specifically, we find that the LLMs which capture syntactic structures better, also have higher abilities in identifying sentence analogies.

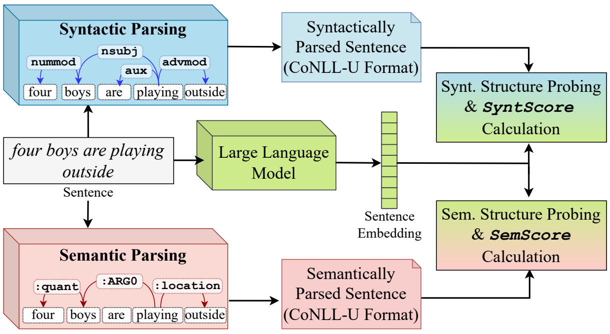

Figure 1: This pipeline details the process of quantifying the LLMs abilities to capture sentence structure via `SyntScore` and `SemScore` values for a given sentence. In this work, we apply this process to a dataset of 100K sentences. The dataset is divided into 0.8 for training the structure probe and 0.1 for testing.

1 Introduction

Analogies facilitate the transfer of meaning and knowledge from one domain to another. Making and identifying analogies is a central tenet in human cognition (Hofstadter, 2001; Holyoak et al., 2001) and is aided by humans' ability to process the structure of language. In the domain of NLP, several types of textual analogies are identified, such as word analogies (Yuan et al., 2023; Gladkova et al., 2016; Gao et al., 2014), proportional word analogies (Chen et al., 2022; Ushio et al., 2021; Szymanski, 2017; Drozd et al., 2016), sentence-analogies (Afantenos et al., 2021; Zhu and de Melo, 2020; Wang and Lepage, 2020) and more recently analogies of procedural/long text (Sultan and Shahaf, 2022). This work explicitly looks at sentence-level analogies which are sentence pairs that are analogues in meaning to each other [1].

Despite the existence of several established benchmarks (e.g., SuperGLUE (Wang et al., 2019a) and GLUE (Wang et al., 2018)) which evaluate the abilities of LLMs extrinsically, Wijesiriwardene et al. (2023) propose a more challenging intrinsic benchmark that focuses on LLMs' ability to identify analogies across a range of complexities.

Identification of analogies relies on the utilization of implicit relational knowledge embedded within the relational structure of language (Gentner, 1983).

In this work we aim to explore the relationship between sentence analogy identification abilities and syntactic and semantic structure encoding abilities of LLMs[2].

Specifically, our main contribution is an analysis of the relationship between the analogy identification ability and sentence structure encoding abilities of LLMs. Additionally, we extend the sentence structure probing techniques introduced by Hewitt and Manning (2019) (which only supports BERT and ELMo) to further work with encoder-decoder-based LLMs and LLMs that use two transformer

*Corresponding author
†Work does not relate to position at Amazon.
[1]For more details on sentence analogies please refer to (Wijesiriwardene et al., 2023)

[2]Our code is available at: https://github.com/Thiliniiw/llms-synt-struct-sentence-analogies

Findings of the Association for Computational Linguistics: EACL 2024, pages 451–457
March 17-22, 2024

architectures. Finally, we extend the structure probing technique originally used for syntactic structure probing in the novel context of semantic structure probing.

2 Related Work

Assessing the ability of Neural Networks (NN) to encode syntactic and semantic structures of language is well examined in NLP (Nivre et al., 2007; Manning and Schutze, 1999; Parsing, 2009). Everaert et al. (2015) emphasize that the meaning of sentences is inferred by the hierarchical structures provided by syntactic and semantic properties of language.

Syntactic parsing aims to derive the syntactic dependencies in a sentence, such as subjects, objects, quantifiers, determiners and other similar elements. Early probing tasks (Adi et al., 2016; Shi et al., 2016) tried to identify NNs' abilities to capture syntactic structures by classifying sentences with single and plural subjects. Later, Conneau et al. (2018) showed that NNs could capture the maximal parse tree depth. The structure probing technique used and extended in this work (Hewitt and Manning, 2019) is related but distinct due to its ability to implicitly capture the parse tree structures through simple distance measures between the vector representations of the words.

Compared to syntactic parsing, the NLP communities' interest in semantic parsing is growing. Semantic parsing maps natural language sentences to a complete, formal meaning representation. Semantic parsing is achieved via combining the Semantic Role Labelling (SRL) approaches with syntactic dependency parsing (Hajic et al., 2009; Surdeanu et al., 2008) and more recently via semantic dependency parsing (Oepen et al., 2014, 2015). This work uses the semantic dependency parsing approach based on mean field variational inference (MFVI) augmented with character and lemma level embeddings introduced by Wang et al. (2019b).

3 Approach

Our approach to exploring the relationship between analogy identification and sentence structure encoding in LLMs is detailed in the following three subsections. We explain the dataset used, in Section 3.1, the analogy identification abilities of LLMs in Section 3.2 and the sentence structure encoding abilities of LLMs in Section 3.3.

Analogy Taxo. Level	Datasets	# Sentences
Level Three	Random deletion/masking/reorder	69,111
Level Four	Negation	1,245
Level Five	Entailment	29,644
Total # Sentences		**100,000**

Table 1: Dataset statistics.

3.1 Dataset

We experiment on a dataset of 100K English sentences. Specifically, the dataset used in this work is randomly picked from the sentence corpus of levels three, four and five of the analogy taxonomy introduced in (Wijesiriwardene et al., 2023). The composition of the dataset is presented in Table 1 (duplicates removed). Specifically, we obtain sentence-analogy pairs provided by Wijesiriwardene et al. (2023) and split the pairs to obtain single sentences used in this work.

3.2 Large Language Models and their Ability to Capture Sentence Analogies

We experiment on the eight language models used in a study by Wijesiriwardene et al. (2023) namely, BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), LinkBERT (Yasunaga et al., 2022), SpanBERT (Joshi et al., 2020) and XLNet (Yang et al., 2019) which are encoder-based LLMs, T5 (Raffel et al., 2020), an encoder-decoder-based LLM and ELECTRA (Clark et al., 2020), an LLM based on two transformer architectures. We refer readers to cited publications for details on the specific LLMs.

Wijesiriwardene et al. (2023) introduced a taxonomy of analogies starting from less complex word-level analogies to more complex paragraph-level analogies and evaluated how each LLM performs on identifying analogies at each level of the taxonomy. An analogy is a pair of lexical items that are identified to hold a similar meaning to each other. Therefore the distance between a pair of analogous lexical items in the vector space should be smaller. The same authors identify Mahalanobis Distance (MD) (Mahalanobis, 1936) to be a better measurement of the distance between two analogous sentences in the vector space. Therefore in this work, the ability of each LLM to identify sentence analogies is represented by the mean MD calculated for the sentence-level datasets (levels 3, 4 and 5) present in the analogy taxonomy. These mean values are calculated based on the reported values by Wijesiriwardene et al. (2023).

3.3 Large Language Models and their Ability to Capture Sentence Structures

Hewitt and Manning (2019) introduced a probing approach to evaluate whether syntax trees (sentence structures) are encoded in Language Models' (LMs') vector geometry. The probing model is trained on train/dev/test splits of the Penn Treebank (Marcus et al., 1993) and tested on both BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018). An LM's ability to capture sentence structure is quantified by its ability to correctly encode the gold parse tree (provided in the Penn Treebank dataset) within its embeddings for a given sentence.

The authors introduce a path distance metric and a path depth metric for evaluation. The distance metric captures the path length between each pair of words measured by Undirected Unlabeled Attachment Score (UUAS) and average Spearman correlation of true to predicted distances (DSpr). The depth metric evaluates the model's ability to identify a sentence's root, measured as root accuracy percentage. Additionally, the depth metric also evaluates the ability of the model to recreate the word order based on their depth in the parse tree identified as Norm Spearman (NSpr.)[3] We refer the readers to Hewitt and Manning (2019) for further details on the technique and evaluation metrics.

4 Experimental Setup

Exploring the relationship between analogy identification and sentence structure encoding abilities of LLMs requires a representative score to quantify (i) analogy identification ability (`AnalogyScore`), (ii) semantic structure identification ability (`SemScore`), and (iii) syntactic structure identification ability (`SyntScore`) of each LLM.

We obtain `AnalogyScore` by calculating the means of reported MD measures obtained for each sentence-level dataset in Wijesiriwardene et al. (2023).

To obtain the `SemScore` (see Figure 1), we first parse all the sentences in our dataset using the MFVI approach (Wang et al., 2019b). The resulting semantically parsed sentences (in CoNLL-U format)[4] and the LLM embeddings of the original sentences are then sent to the structure probing technique (Hewitt and Manning, 2019). The structure probe is trained on 80K sentences from the dataset and the DSpr and UUAS values representing parse distance and root accuracy (RootAcc) value representing parse depth are reported on the test split with 10K sentences. Finally, the `SemScore` is computed as a combined score by taking the mean of the z-score normalizations of these three measures $Z_{DSpr}, Z_{UUAS}, Z_{RootAcc}$ (see Table 2).

$$\texttt{SemScore} = \frac{1}{3}(Z_{DSpr} + Z_{UUAS} + Z_{RootAcc})$$

$$\texttt{SyntScore} = \frac{1}{3}(Z_{DSpr} + Z_{UUAS} + Z_{RootAcc})$$

To obtain the `SyntScore` (see Figure 1), we follow the same steps but parse the sentences syntactically. Finally, we calculate the Spearman's rank correlation (SRC) and Kendall's rank correlation (KRC) between `AnalogyScore` and `SyntScore`, as well as `AnalogyScore` and `SemScore`.

4.1 Implementation Details

When extending the structure probing technique by Hewitt and Manning (2019) to facilitate additional LLMs, we use the HuggingFace implementation[5] of the LLMs. For semantic parsing, we use the trained mean field variational inference (MFVI) model augmented with character and lemma-level embeddings provided by the SuPar[6]. For syntactic parsing of the sentences we employ Stanford CoNLL-U dependency parser[7].

5 Results

In this section, we look at the findings of this work with regard to semantic and syntactic structure encoding abilities and analogy identification abilities of LLMs.

5.1 Semantic and Syntactic Structure Encoding Abilities of LLMs

We tabulate the structure probing results in original metrics (Table 2) and the performance of each LLM in identifying sentence analogies and capturing the semantic and syntactic structures (Table 3). It is interesting to note that RoBERTa, the best-performing LLM for analogy identification (`AnalogyScore` $= 0.458$), holds the highest `SyntScore` and `SemScore`. XLNet is the lowest-performing model for analogy identification as well as syntactic structure identification. Yet it performs second-best in semantic structure identification. SpanBERT ranks second in both analogy

[3] We do not use NSpr. in this work.

[4] `https://universaldependencies.org/format.html`

[5] `https://huggingface.co/models`

[6] `https://github.com/yzhangcs/parser`

[7] `https://nlp.stanford.edu/software/nndep.html`

Model	Original Scores						Normalized Scores					
	Syntactic			Semantic			Syntactic			Semantic		
	Distance		Depth	Distance		Depth	Distance		Depth	Distance		Depth
	DSpr	UUAS	RootAcc	DSpr	UUAS	RootAcc	Z_{DSpr}	Z_{UUAS}	$Z_{RootAccu}$	Z_{DSpr}	Z_{UUAS}	$Z_{RootAccu}$
ALBERT	0.59	0.46	0.35	0.38	0.13	0.19	-1.56	-2.30	-2.58	0.39	-1.30	0.36
BERT	0.73	0.72	0.74	0.38	0.16	0.17	0.87	0.62	0.56	0.39	-0.03	0.07
Electra	0.70	0.76	0.75	0.38	0.14	0.15	0.34	1.01	0.63	0.39	-0.73	-0.28
LinkBERT	0.70	0.68	0.69	0.38	0.15	0.05	0.33	0.18	0.15	0.37	-0.27	-1.79
RoBERTa	0.74	0.74	0.73	0.38	0.16	0.29	1.06	0.77	0.49	0.37	0.25	1.89
SpanBERT	0.74	0.72	0.74	0.38	0.14	0.20	1.06	0.56	0.55	0.37	-0.97	0.54
T5	0.63	0.64	0.71	0.37	0.19	0.17	-0.79	-0.31	0.28	-2.65	1.64	0.05
XLNet	0.60	0.62	0.66	0.38	0.18	0.11	-1.31	-0.53	-0.08	0.37	1.42	-0.83

Table 2: DSpr, UUAS measures indicating Parse Distance (Distance) and RootAcc measure indicating Parse Depth (Depth). Original Scores denote original output values of the structure probe technique and Normalized Scores are z-score normalized. Higher values indicate a stronger ability of the LLMs to capture sentence structures.

Model	AnalogyScore		SyntScore		SemScore	
	Score	Rank	Score	Rank	Score	Rank
ALBERT	0.645	7	-2.14	8	-0.19	5
BERT	0.505	3	0.68	3	0.14	3
Electra	0.516	4	0.66	4	-0.21	6
LinkBERT	0.608	6	0.22	5	-0.56	8
RoBERTa	**0.458**	**1**	**0.78**	**1**	**0.84**	**1**
SpanBERT	0.461	2	0.72	2	-0.02	4
T5	0.524	5	-0.27	6	-0.32	7
XLNet	0.747	8	-0.64	7	0.32	2

Table 3: The values for `AnalogyScore`, `SyntScore` and `SemScore` and their corresponding rank values. `AnalogyScore` ranges between [0,1], 0 being the best. For `SyntScore` and `SemScore` higher the values better the ability of LLMs to capture sentence structure.

identification and syntactic structure identification but holds the median `SemScore`.

5.2 Analogy Identification and Syntactic Structure Encoding Abilities of LLMs

We use SRC and KRC values to analyze the correlation between LLMs' ability to identify sentence analogies denoted by `AnalogyScore` and LLMs' ability to encode syntactic structures of sentences denoted by `SyntScore`. Both correlation measures show a significant positive correlation between `AnalogyScore` and `SyntScore`. Specifically, the SRC between `AnalogyScore` and `SyntScore` is 0.95 ($p < 0.001$). The KRC between `AnalogyScore` and `SyntScore` is 0.86 ($p = 0.002$).

5.3 Analogy Identification and Semantic Structure Encoding abilities of LLMs

Similar to the previous section, we compute the SRC and KRC values to asses the correlations between `AnalogyScore` and `SemScore`. We see that both correlations are positive with SRC of 0.33 ($p = 0.42$) and KRC of 0.28 ($p = 0.40$) between `AnalogyScore` and `SemScore`.

6 Limitations

Several contemporary probing techniques, such as those outlined in Voita and Titov (2020) and Pimentel et al. (2020), have emerged subsequent to the methodology employed in the present investigation (Hewitt and Manning, 2019). Nevertheless, in the context of our current study, we have only chosen to employ (Hewitt and Manning, 2019) owing to its adaptable nature, which facilitates extension to various LLMs that are of particular interest to our current research.

Even though Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a popular and widely used technique to parse sentences semantically, in current work, we use MFVI, a semantic parsing approach introduced by Wang et al. (2019b) because of the limitations posed by the structure probing technique used (Hewitt and Manning, 2019). This technique requires the mapped LLM embeddings and semantic dependency parsed sentences to be of the same length. However, as it is known, AMRs abstract away from the syntactic idiosyncrasies of the language and overlook certain auxiliary words from the parse results, limiting its use in this work.

The present study employs a semantic parsing technique reported to exhibit a high accuracy level of 94% (Wang et al., 2019b). However, it is important to note that for the purposes of our investigation, we make the assumption that the semantically parsed sentences generated by this particular method are entirely accurate, thereby employing them as the gold standard data. It is worth mentioning that this choice may introduce some degree of bias into our examination of the semantic structure probing.

7 Conclusion and Future Directions

This work explores the relationship between LLMs' ability to identify sentence analogies and encode sentence structures in their embeddings. Through detailed experiments, we show that the sentence analogy identification ability of LLMs is positively correlated with their ability to encode syntactic and semantic structures of sentences. Particularly, LLMs that better capture syntactic structures have a higher correlation to analogy identification. In summary this work explores how LLMS utilize the knowledge of semantic and syntactic structures of sentences to identify analogies. Moving forward, we aim to explore the potential of extending the current approach to enhance explainability of LLMs within the broader domain of NLP.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work was supported in part by the NSF grant #2335967: EAGER: Knowledge-guided neurosymbolic AI with guardrails for safe virtual health assistants. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organization.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Stergos Afantenos, Tarek Kunze, Suryani Lim, Henri Prade, and Gilles Richard. 2021. Analogies between sentences: Theoretical aspects-preliminary experiments. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21–24, 2021, Proceedings 16*, pages 3–18. Springer.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-kar: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.

Martin BH Everaert, Marinus AC Huybregts, Noam Chomsky, Robert C Berwick, and Johan J Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19(12):729–743.

Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. Wordrep: A benchmark for research on learning word representations. *arXiv preprint arXiv:1407.1640*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Douglas R Hofstadter. 2001. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538.

Keith J Holyoak, Dedre Gentner, Boicho N Kokinov, and Franz Hall. 2001. The place of analogy in cognition.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statistics. National Institute of Science of India.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.

Constituency Parsing. 2009. Speech and language processing. *Power Point Slides*.

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.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*, pages 448–453.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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.

Liyan Wang and Yves Lepage. 2020. Vector-to-sequence models for sentence analogies. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 441–446. IEEE.

Xinyu Wang, Jingxian Huang, and Kewei Tu. 2019b. Second-order semantic dependency parsing with end-to-end neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4618.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

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.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. *arXiv preprint arXiv:2305.05994*.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.