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, USA

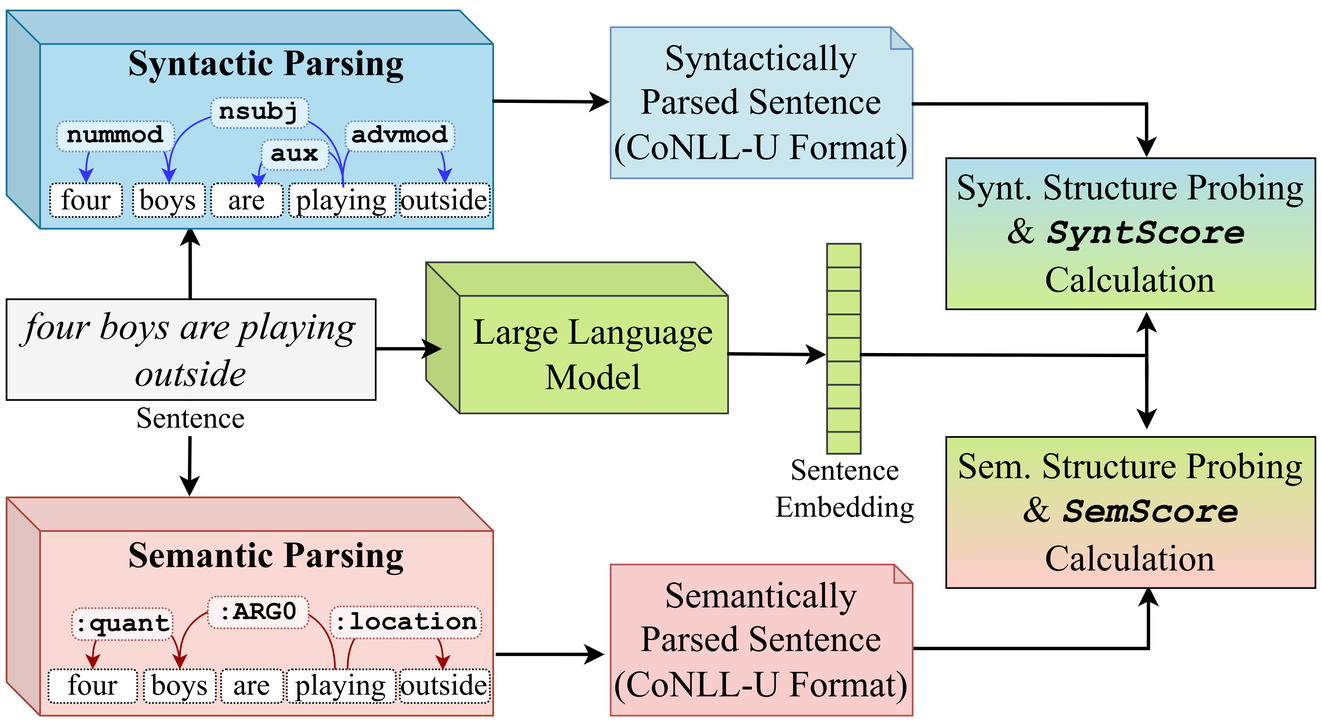

Figure 1: This pipeline details the process of quantifying the LLMs abilities to capture sentence structure via SyntScore and SemScore values for a given sentence.

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 1: 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.

1. Background

- 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.

2. Method

- 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`),
  (iii) syntactic structure identification ability (`SyntScore`)
  of each LLM.
- `AnalogyScore` → means of reported MD measures obtained for each sentence-level dataset in [1].
- `SemScore` (see Figure 1) → parse all the sentences in our dataset using the MFVI approach [2]. The resulting semantically parsed sentences (in CoNLL-U format) and the LLM embeddings of the original sentences are then sent for structure probing [3].
- Structure probe → trained on 80K sentences from the dataset and the Spearman correlation of true to predicted distances (DSpr) and Undirected Unlabeled Attachment Score (UUAS) values representing parse distance and root accuracy (RootAcc) value representing parse depth are reported on the test split with 10K sentences.
- `SemScore & SyntScore` values are computed as a combined score by taking the mean of the z-score normalizations of above three measures (See below and Figure 1).

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

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

3. Findings

- LLMs' ability to identify sentence analogies is positively correlated with their ability to encode syntactic and semantic structures of sentences.
- Specifically, LLMs which capture syntactic structures better, also have higher abilities in identifying sentence analogies.
- `AnalogyScore` & `SyntScore` → Spearman's rank correlation (SRC) is 0.95 ($p < 0.001$). Kendall's rank correlation (KRC) is 0.86 ($p = 0.002$) (See Table 1).
- `AnalogyScore` & `SemScore` → SRC of 0.33 ($p = 0.42$) and KRC of 0.28 ($p = 0.40$) (See Table 1).

4. Limitations

- Only used Hewitt and Manning 2019 probing technique.
- AMR vs. MFVI.
- The present study employs a semantic parsing technique reported to exhibit a high accuracy level of 94% but we assume that the semantically parsed sentences generated by this method are entirely accurate.

References

1. 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.
2. Xinyu Wang, Jingxian Huang, and Kewei Tu. 2019. 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, Florence, Italy. Association for Computational Linguistics.
3. 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.

*Work does not relate to position at Amazon.

Acknowledgement: National Science Foundation (Grant #2335967)

For more information or feedback, Contact Thilini at thilini@sc.edu