Fine-Tuning vs. Prompting : Évaluation de la construction de graphes de connaissances avec les LLMs

Hussam Ghanem*, Christophe Cruz *

*ICB, UMR 6306, CNRS, Université de Bourgogne, 21000 Dijon, France
{prenom.nom}@u-bourgogne.fr

1 Introduction, Contexte et Travaux Connexes

Les Graphes de Connaissances (Knowledge Graphs ou KGs) constituent aujourd'hui un pilier fondamental pour de nombreuses applications avancées. Néanmoins, leur construction à partir de données textuelles non structurées demeure un défi majeur. Les récents progrès des grands modèles de langage (LLMs) ouvrent de nouvelles opportunités pour la construction de KGs. Cet article présente une synthèse de notre publication (Ghanem et Cruz, 2024) traitant des forces et des faiblesses des trois méthodes de construction de KG basées sur des LLMs : Zero-Shot Prompting (ZSP), Few-Shot Prompting (FSP) et Fine-Tuning (FT). Les techniques traditionnelles de construction de KG reposent souvent sur des pipelines séquentiels d'extraction d'information. Les approches de bout en bout présentent un potentiel d'amélioration. Les LLMs tels que GPT-3 [1], LLaMA [2], Mistral [3] et Starling [4] offrent des capacités accrues pour la construction de KGs à partir de texte (T2KG), exploitables via ZSP, FSP et FT. Les travaux précédents ont utilisé des métriques telles que Triple Match F1 (T-F1), Graph Match F1 (G-F1) et Graph Edit Distance (GED) pour l'évaluation (Han et al., 2023), mais négligent souvent les taux d'hallucination et d'omission.

2 Méthodologie, Résultats et Discussion

Nous avons mené des expérimentations avec le jeu de données WebNLG+2020. Trois LLMs (Llama 2 7B, Mistral, Starling) on fait l'objet de notre évaluation. Chaque modèle a été testé avec ZSP, FSP (7 exemples) et FT. Les métriques d'évaluation incluent T-F1, G-F1, G-BS, GED, BLEU-F1 [5], ROUGE-F1 [6] et nous utilisons Optimal Edit Paths (OEP) [7], pour quantifier les hallucinations et omissions. Le tableau 1 résume les résultats des différents LLMs et méthodologies sur le jeu de données WebNLG+2020. Les résultats montrent que l'affinage (FT)

1. Language Models are Few-Shot Learners : https://arxiv.org/abs/2005.14165
2. LLaMA : Open and Efficient Foundation Language Models : https://arxiv.org/abs/2302.13971
3. Mistral 7B : https://arxiv.org/abs/2310.06825
4. Starling-7b : https://starling.cs.berkeley.edu/
5. BLEU : https://aclanthology.org/P02-1040.pdf
6. ROUGE : https://aclanthology.org/W04-1013/
7. NetworkX - optimal edit paths : https://networkx.org/documentation/stable/index.html

TAB. 1 – *Résultats de différents modèles sur WebNLG+2020. Des valeurs plus faibles indiquent de meilleures performances pour GED, Hall. et Omis.*

Modèle | Metrique	G-F1	T-F1	G-BS	GED	F1-Bleu	F1-Rouge	Hall.	Omis.
PiVE	14.00	18.57	89.82	11.22	-	-	-	-
Mistral-0	2.30	0.00	77.87	15.93	54.97	55.15	20.63	31.48
Mistral-7	18.72	28.44	87.54	10.13	55.09	63.94	17.88	21.14
Mistral-FT-0	31.93	44.08	86.89	8.25	63.88	69.08	13.55	18.27
Mistral-FT-7	34.68	**49.11**	**91.99**	6.69	**71.78**	**77.43**	15.01	**14.45**
Starling-0	5.23	7.83	86.29	13.35	34.64	14.61	17.48	33.24
Starling-7	21.30	33.77	90.41	8.96	60.47	69.34	17.31	14.61
Starling-FT-0	21.47	28.29	72.86	11.87	44.07	47.69	10.17	42.78
Starling-FT-7	**35.69**	48.49	91.95	**6.60**	71.51	76.67	11.35	18.27
Llama2-7b-0	0.00	0.46	54.20	18.29	20.23	17.98	4.83	81.53
Llama2-7b-7	11.80	20.88	82.78	12.66	45.48	54.29	20.74	30.02
Llama2-7b-FT-0	3.82	15.41	59.19	15.78	16.82	17.95	6.07	79.20
Llama2-7b-FT-7	18.77	32.63	87.19	10.16	58.48	66.35	25.24	18.66

TAB. 2 – *Résultats sur KELM-sub.*

Modèle | Métrique	G-F1	T-F1	G-BS	GED	F1-Bleu	F1-Rouge	Hall.	Omis.
PiVE	**23.11**	7.50	**87.70**	**11.35**	-	-	-	-
Mistral-7	5.61	10.89	71.29	14.28	56.56	61.11	2.33	77.33
Mistral-FT-7	2.83	8.73	68.55	14.54	26.35	38.76	1.78	78.17
Starling-7	5.61	**13.82**	83.16	12.85	**65.79**	**71.20**	5.33	**59.44**
Starling-FT-7	3.11	9.82	67.79	14.53	27.37	39.49	1.22	78.67

surpasse généralement ZSP et FSP, avec Mistral et Starling (affinés et 7 exemples) affichant de meilleures performances que Llama 2, et que FT est mieux que ZSP et FSP sur les modèles originaux (sans FT). L'augmentation du nombre d'exemples dans FSP améliore généralement les performances. Les modèles affinés surpassent nettement une base précédente (PiVE). Cependant, sur le sous-ensemble KELM-sub (tableau 2), les modèles non affinés avec 7 exemples surpassent parfois les modèles affinés, suggérant des défis de généralisation qui méritent une exploration approfondie. Ainsi, de travaux futurs devraient se concentrer sur le raffinement des métriques pour gérer les synonymes et exploiter les LLMs pour l'augmentation de données afin d'améliorer la généralisation.

Références

Ghanem, H. et C. Cruz (2024). Fine-tuning vs. prompting : Evaluating the knowledge graph construction with llms. *International Workshop On Knowledge Graph Generation From Text (TEXT2KG), Co-located with the Extended Semantic Web Conference (ESWC), https ://ceur-ws.org/Vol-3747/text2kg_paper7.pdf* .

Han, J., N. Collier, W. Buntine, et E. Shareghi (2023). Pive : Prompting with iterative verification improving graph-based generative capability of llms. *arXiv preprint arXiv :2305.12392*.