Conversational Ontology Alignment with ChatGPT

Sanaz Saki Norouzi[1], Mohammad Saeid Mahdavinejad[1] and Pascal Hitzler[1]

[1]*Department of Computer Science, Kansas State University, USA*

Abstract

This study evaluates the applicability and efficiency of ChatGPT for ontology alignment using a naive approach. ChatGPT's output is compared to the results of the Ontology Alignment Evaluation Initiative 2022 campaign using conference track ontologies. This comparison is intended to provide insights into the capabilities of a conversational large language model when used in a naive way for ontology matching and to investigate the potential advantages and disadvantages of this approach.

Keywords

Ontology alignment, ChatGPT, Schema matching, Ontology matching, Large language models, Prompt engineering, LLM behavior

1. Introduction

Ontology alignment (OA), also referred to as ontology matching, is a central task in semantic web technologies that aims to find semantic correspondences between two ontologies with overlapping domains. As using ontologies is extending to many different fields, this task's importance is increasing, so ontology matching is required for bridging the semantic gap between various ontologies [1]. Although OA already looks back to many years of research, the task remains challenging, often requiring expert intervention to ensure accurate results. Expert-driven matching can be both time-consuming and subject to human biases, so even in this case absolute precision remains elusive [2, 3, 4]. To tackle this challenge, a variety of ontology matching systems, incorporating natural language processing (NLP) techniques considering grammar changes and different similarity measurements, machine learning, fuzzy lexical matching, and other advanced methodologies are proposed in the Ontology Alignment Evaluation Initiative (OAEI) 2022 [5]. Each approach attempts to automate the matching process and alleviate the need for extensive human involvement.

With the emergence of large language models (LLMs), we have seen impressive results in many NLP downstream tasks. Recently, using LLMs is increased for human-centric tasks, and models like ChatGPT[1] by OpenAI[2] have attracted attention for doing different tasks such as logical reasoning [6], question answering [7], and mental health analysis [8]. Prompt engineering is a skill that is required to work with LLMs efficiently. A prompt can be considered as a direction to interact with LLMs to adjust and control their output [9]. Generally, for using LLMs, there

OM 2023: The 18th International Workshop on Ontology Matching collocated with the 22nd International Semantic Web Conference ISWC-2023 November 7th, 2023, Athens, Greece

sanazsn@ksu.edu (S. Saki Norouzi); saeid@ksu.edu (M. S. Mahdavinejad); hitzler@ksu.edu (P. Hitzler)

[1]ChatGPT refers to ChatGPT version 4.0 unless otherwise specified.

[2]https://chat.openai.com/chat

are three main approaches: fine-tuning, few-shot prompting, and zero-shot prompting. For using some LLMs in downstream tasks, fine-tuning would be helpful since it would make the LLM adapt its knowledge (from the pre-training process) to the specific task. Recently, as it is reported, models like GPT-3 [10] are able to generate responses to some tasks that it has not been trained on, so prompt engineering became more popular. In few-shot prompting, a few examples of the task and the format of input/output are given to the model, so it would be able to give the output based on the format while in zero-shot prompting it is only possible to evaluate the performance of the LLM based on its knowledge in one prompt. Thus, prompt patterns are important in the results provided by these LLMs.

In this paper, we conduct a comparative analysis of ChatGPT's performance in ontology alignment when prompted with different strategies. We compare ChatGPT's output with the reference alignments provided by the Ontology Alignment Evaluation Initiative (OAEI) 2022 campaign, which uses conference-related ontologies. By evaluating ChatGPT's performance in a zero-shot manner, we aim to shed light on the capabilities and limitations of using a conversational large language model for ontology matching. Furthermore, we discuss the implications of our findings and propose potential directions for future research in this exciting area.

2. Methodology

Data

Our evaluation focuses on conference track ontologies provided by the OAEI [11], encompassing seven ontologies: cmt, conference, sigkdd, iasted, ekaw, edas, and confOf. This selection yields 21 pairs of matched ontologies. We use the original reference alignment known as ra1[3] for our evaluation. It is mentioned by OAEI, that M3 evaluation means both properties and classes are considered for matching. Thus, we consider ra1-M3 OAEI 2022 results for comparison.

Prompts and Formatting

An essential aspect of this evaluation involves designing prompts that effectively incorporate the triples from the conference track ontologies. We explore different approaches to include ontology triples in the prompts, with two primary methods considered: converting triples into sentences and transforming them into formatted text following the pattern Predicate(Subject, Object).

After conducting experiments and considering the effectiveness of different prompt approaches, we choose to adopt the formatted text approach for our prompts, which aligns well with suggestions from OpenAI. This formatting presents triples in a structured manner, making it easier for ChatGPT to comprehend and generate appropriate responses. For instance, an original triple such as "track subclassOf conference_part" can be represented as "Is-a (track, conference part)" using the formatted text approach. Similarly, properties are expressed in the same structured format, such as "authorOf (Person, Document)".

[3]https://oaei.ontologymatching.org/2023/conference/data

The limitation of a basic version of ChatGPT (v3.5), which we will elaborate on more in the discussion section, led us to divide it into smaller parts instead of using one long prompt. This approach allowed us to maintain essential context throughout the interaction, resulting in a better understanding of the model and more accurate responses.

In our early experiments, we found that adding more complex ontology axioms made it more difficult for ChatGPT to capture the best possible matches between two ontologies. Therefore, we decided to include only axioms that can be directly expressed as triples. We formulated our prompt with a structured approach as follows:

<Problem Definition>
In this task, we are given two ontologies in the form of Relation(Subject, Object), which consist of classes and properties.
<Ontologies Triples>
Ontology 1:
Ontology 1 Triples
Ontology 2:
Ontology 2 Triples
<Objective>
Our objective is to provide ontology mapping for the provided ontologies based on their semantic similarities.

Table 1 illustrates the diverse set prompt designs and formatting approaches used to assess ChatGPT's ontology alignment effectiveness.

3. Results and Analysis

In this section, we present the results of our evaluation. The objective was to gain insights and investigate this approach's potential advantages and disadvantages. Among the prompts, "prompt 7" demonstrated the highest recall. However, it should be noted that the number of generated statements for this prompt was relatively higher than "prompt 1" since it is a repetitive prompt for each class/property name, and it tries to find the best match for each of them. Thus, the increased recall came at the cost of reduced precision, while it should be noted that some of the generated statements were deemed irrelevant even by non-expert evaluators. Nonetheless, "prompt 7" exhibited the highest F1-score among all the prompts, showcasing a balance between recall and precision.

While the first three prompts are similar in essence but have different objectives, their F1-scores are almost the same. Asking for a complete and comprehensive matching gives the highest recall, but also the least precision. On average, the first prompt achieved the best balance between recall and precision. Interestingly, employing prompts that explicitly asked for matching classes or properties, such as prompts 4 and 5, resulted in higher recall but lower precision and F1-scores. Nevertheless, this drawback can be mitigated by domain experts who can easily filter out irrelevant generated statements. For a more comprehensive evaluation, we compare our results with OAEI 2022 results in Table 2. The prompts' results are shown in Table 3.

Table 1
Details of the prompts in each experiment. P# shows the prompt number.

P#	Description	Prompt structure
1	Put all the information in a single prompt.	<Problem Definition> <Ontologies Triples> <Objective>
2	Changing the objective of the prompts.	<Problem Definition> <Ontologies Triples>
		Provide a complete and comprehensive matching of the ontologies
3	Changing the objective of the prompt.	<Problem Definition> <Ontologies Triples>
		Match these two ontologies and provide the most accurate matching you can do
4	Separate the class and data/object properties in two consecutive prompts.	<Problem Definition> <Class Triples>
		<Data/Object Triples> <Objective>
5	Following the Exp 2 pattern but changing the objective of the prompt.	<Problem Definition> <Class Triples>
		<Data/Object Triples> Match these two ontologies and provide the most accurate matching you can do
6	Following the Exp 2 pattern but changing the order of triples to prioritizing the root class entities.	<Problem Definition> <Class Triples>
		<Data/Object Triples> <Objective>
7	First, Providing the Ontologies, then asks about the best class/property of the second ontology that can be matched with the class/property of the first one.	<Problem Definition> <Ontologies Triples>
		For a class/property in the first ontology, which class/property in ontology 2 is the best match? <Ontology 2 Triples>

Table 2
Comparison of OAEI 2022 results with ChatGPT

Matcher	Precision	Recall	F1-score
ALIN	**0.88**	0.47	0.61
ALIOn	0.75	0.22	0.34
AMD	0.87	0.43	0.58
ATMatcher	0.74	0.53	0.62
edna	0.79	0.47	0.59
GraphMatcher	0.8	0.57	0.67
KGMatcher+	**0.88**	0.4	0.55
LogMap	0.81	0.58	**0.68**
LogMapLt	0.73	0.5	0.59
LSMatch	**0.88**	0.42	0.57
Matcha	0.38	0.08	0.13
SEBMatcher	0.84	0.5	0.63
StringEquiv	0.8	0.43	0.56
TOMATO	0.09	0.63	0.16
ChatGPT-4	0.37	**0.92**	0.52

4. Discussion

Our evaluation highlighted a significant challenge related to precision. The generated statements often introduced errors that caused a decrease in precision. We identified several factors contributing to this issue:

ChatGPT context length limit: ChatGPT (v4.0) was used in our experiments because ChatGPT (v3.5) struggled to retain context when the input was lengthy, affecting its performance in ontology alignment tasks. ChatGPT (v4.0) has improved contextual understanding and better adaptability to long inputs, and its maximum token length of 8192 accommodates both ontology triples within the prompt.

Inverse Functional Properties: These Properties can lead to imprecise matches if they are not properly accounted for. For example, the statement hasBeenAssigned(Reviewer, Paper) is matched to hasReviewer(Paper, Possible_Reviewer) by ChatGPT. However, the correct entity for this matching is ReviewerOfPaper, which is the inverse of hasReviewer. If we properly account for this inverse relationship, we can enhance precision by reducing the number of false positives.

Matches with Subclasses: The generated alignments sometimes matched a class in one ontology to one class and all its subclasses in the other, leading to unintended matches. For instance in the conference-edas matching, "active_conference_participant" and "passive_conference_participant" which are subclasses of conf_participant are matched with at-

Table 3
Comparison of precision, recall, and F1-score for different prompts. The cells marked with a dash (-) couldn't be completed due to token input limitations. P, R, F1 show precision, recall, and F1-score, respectively.

Dataset	prompt 1			prompt 2			prompt 3			prompt 4			prompt 5			prompt 6			prompt 7		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
cmt-conference	0.437	0.466	0.45	0.28	0.466	0.35	0.5	0.466	0.48	0.275	0.533	0.36	0.4	0.8	0.53	0.478	0.733	0.58	0.304	0.933	0.46
cmt-ekaw	0.533	0.727	0.61	0.5	0.727	0.59	0.388	0.636	0.48	0.321	0.818	0.46	0.21	0.727	0.33	0.346	0.818	0.49	0.26	0.909	0.40
cmt-iasted	-	-	-	-	-	-	-	-	-	0.173	1	0.29	0.266	1	0.42	0.25	0.75	0.37	0.072	1	0.13
cmt-sigkdd	1	0.666	0.8	0.363	0.666	0.47	0.5	0.666	0.57	0.75	1	0.86	0.625	0.833	0.71	0.75	0.75	0.75	0.461	1	0.63
cmt-confOf	0.538	0.437	0.48	0.36	0.562	0.44	0.833	0.312	0.45	0.47	0.562	0.51	0.4	0.5	0.44	0.4	0.625	0.49	0.411	0.875	0.56
cmt-edas	0.666	0.615	0.64	0.769	0.769	0.77	0.562	0.692	0.62	0.529	0.692	0.6	0.354	0.846	0.5	0.346	0.692	0.46	0.28	0.923	0.43
conference-ekaw	0.411	0.28	0.33	0.55	0.44	0.49	0.52	0.48	0.5	0.333	0.52	0.41	0.344	0.4	0.37	0.25	0.48	0.33	0.38	0.92	0.54
conference-iasted	-	-	-	-	-	-	-	-	-	0.285	0.428	0.34	0.277	0.357	0.31	0.208	0.357	0.26	0.325	0.928	0.48
conference-sigkdd	0.6	0.4	0.48	0.379	0.733	0.5	0.45	0.6	0.51	0.413	0.8	0.54	0.232	0.666	0.34	0.26	0.4	0.31	0.407	0.733	0.52
conference-confOf	0.35	0.466	0.40	0.222	0.666	0.33	0.4	0.533	0.46	0.357	0.666	0.46	0.307	0.533	0.39	0.366	0.733	0.49	0.466	0.933	0.62
conference-edas	0.28	0.411	0.33	0.45	0.529	0.49	0.529	0.529	0.53	0.375	0.529	0.44	0.257	0.529	0.34	0.323	0.647	0.43	0.35	0.882	0.50
ekaw-iasted	-	-	-	-	-	-	-	-	-	0.352	0.6	0.44	0.222	0.4	0.28	0.181	0.2	0.19	0.322	1	0.49
ekaw-sigkdd	0.466	0.636	0.54	0.36	0.818	0.5	0.411	0.636	0.5	0.28	0.636	0.39	0.454	0.909	0.60	0.666	0.727	0.69	0.33	1	0.67
confOf-ekaw	0.5	0.75	0.6	0.478	0.55	0.51	0.518	0.7	0.59	0.355	0.8	0.49	0.448	0.65	0.53	0.625	0.75	0.68	0.558	0.95	0.70
confOf-sigkdd	0.19	0.571	0.28	0.357	0.714	0.48	0.235	0.571	0.33	0.181	0.571	0.27	0.23	0.857	0.36	0.357	0.714	0.48	0.318	1	0.48
confOf-edas	0.428	0.631	0.51	0.454	0.526	0.49	0.428	0.631	0.51	0.363	0.631	0.46	0.425	0.894	0.58	0.545	0.631	0.58	0.444	0.842	0.58
confOf-iasted	0.555	0.555	0.55	0.461	0.666	0.54	0.466	0.777	0.58	0.266	0.444	0.33	0.347	0.888	0.5	0.206	0.666	0.31	0.241	0.777	0.37
edas-ekaw	0.6	0.391	0.47	0.423	0.478	0.45	0.588	0.434	0.5	0.55	0.478	0.51	0.484	0.695	0.57	0.464	0.565	0.51	0.466	0.913	0.62
edas-iasted	-	-	-	-	-	-	-	-	-	0.384	0.263	0.31	0.352	0.631	0.45	0.307	0.210	0.25	0.38	0.842	0.52
edas-sigkdd	0.5	0.333	0.4	0.555	0.666	0.6	0.647	0.733	0.69	0.473	0.6	0.53	0.608	0.933	0.74	0.444	0.8	0.57	0.535	1	0.7
iasted-sigkdd	0.75	0.6	0.67	0.4	0.266	0.32	0.384	0.333	0.36	0.370	0.666	0.48	0.466	0.466	0.47	0.4	0.666	0.5	0.384	1	0.55
Average	0.52	0.52	0.50	0.43	0.60	0.49	0.49	0.57	0.51	0.37	0.63	0.45	0.37	0.69	0.46	0.39	0.61	0.46	**0.37**	**0.92**	**0.52**

tendee from the other ontology. Addressing this scenario is crucial for improving alignment accuracy.

Unseen/Ambiguous Alignments: Some generated alignments may appear to be accurate to non-experts, but they are actually incorrect according to reference datasets. This will be a challenge for LLMs. To address this issue, we propose two possible solutions: (1) revising the reference datasets to eliminate any ambiguity or inconsistency in the alignment criteria, or (2) developing a method to help LLMs detect and avoid generating implausible alignments. For instance, "camera_ready_paper" and "final_manuscript" seem similar.

Uncertain Matching: In certain cases, even though ChatGPT acknowledges that a matching is unlikely, it still generates such matches and proposes new entities to be included in the graph.

5. Conclusion and Future Work

In this paper, we have evaluated the applicability and efficiency of ChatGPT for ontology alignment using a naive approach. Our evaluation showed that ChatGPT can achieve high recall but also suffers from low precision. We identified several factors contributing to this issue, including the context length limit of ChatGPT, the handling of inverse functional properties, the matching with subclasses, unseen alignments, and uncertain matchings. Despite the mentioned challenges, we believe that ChatGPT has the potential to be a valuable tool for ontology alignment. The high recall of ChatGPT means that it can be used to identify a large number of potential matches, which can then be filtered by domain experts. Additionally, the ability of ChatGPT to generate new entities suggests that it could be used to expand reference ontologies. In future work, we plan to address the precision issues identified in this paper. We also plan to explore other ways to use ChatGPT for ontology alignment, such as generating prompts

for more sophisticated alignment algorithms. Overall, we believe that the results of this paper demonstrate the potential of ChatGPT for ontology alignment. We believe that this approach can be used to improve the efficiency and effectiveness of ontology alignment tasks.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant 2033521 A1. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] P. Shvaiko, J. Euzenat, Ontology matching: state of the art and future challenges, IEEE Transactions on knowledge and data engineering 25 (2011) 158–176.

[2] C. Trojahn, R. Vieira, D. Schmidt, A. Pease, G. Guizzardi, Foundational ontologies meet ontology matching: A survey, Semantic Web 13 (2022) 685–704.

[3] R. Stevens, P. Lord, J. Malone, N. Matentzoglu, Measuring expert performance at manually classifying domain entities under upper ontology classes, Journal of Web Semantics 57 (2019) 100469.

[4] M. Cheatham, P. Hitzler, Conference v2.0: An uncertain version of the OAEI conference benchmark, in: P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. Groth, N. F. Noy, K. Janowicz, C. A. Goble (Eds.), The Semantic Web – ISWC 2014 – 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II, volume 8797 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 33–48. doi:10.1007/978-3-319-11915-1_3.

[5] M. A. N. Pour, A. Algergawy, P. Buche, L. J. Castro, J. Chen, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, et al., Results of the ontology alignment evaluation initiative 2022, CEUR Workshop Proceedings, 2023.

[6] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, Y. Zhang, Evaluating the logical reasoning ability of chatgpt and gpt-4, arXiv preprint arXiv:2304.03439 (2023).

[7] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, G. Qi, Evaluation of chatgpt as a question answering system for answering complex questions, arXiv preprint arXiv:2303.07992 (2023).

[8] K. Yang, S. Ji, T. Zhang, Q. Xie, S. Ananiadou, On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis, arXiv preprint arXiv:2304.03347 (2023).

[9] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[11] O. Zamazal, V. Svátek, The ten-year ontofarm and its fertilization within the onto-sphere, Journal of Web Semantics 43 (2017) 46–53.