



**HAL**  
open science

# Automatic Ontology Population from Textual Advertisements: LLM vs. Semantic Approach

Aya Nour Elimane Sahbi, Céline Alec, Pierre Beust

## ► To cite this version:

Aya Nour Elimane Sahbi, Céline Alec, Pierre Beust. Automatic Ontology Population from Textual Advertisements: LLM vs. Semantic Approach. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES), Sep 2024, Séville, Espagne, France. pp.3083-3092, <10.1016/j.procs.2024.09.364>. <hal-04814520>

**HAL Id: hal-04814520**

**<https://hal.science/hal-04814520v1>**

Submitted on 22 May 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License



28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

# Automatic Ontology Population from Textual Advertisements: LLM vs. Semantic Approach

Aya Sahbi<sup>a,\*</sup>, Céline Alec<sup>a</sup>, Pierre Beust<sup>a,b</sup>

<sup>a</sup>Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

<sup>b</sup>Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France

## Abstract

Automatic ontology population involves identifying, extracting and integrating information from various sources to instantiate the classes and properties of an ontology, thereby building a domain Knowledge Graph (KG). In this paper, we compare two text-based ontology population techniques: KOnPoTe, a semantic approach based on textual and domain knowledge analysis, and a generative AI approach utilizing Claude, a Large Language Model (LLM). We present experiments conducted on two French sales advertisement domains: real estate and boats, and discuss the strengths and limitations of both approaches.

© 2024 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0>)

Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems

**Keywords:** Ontology Population; LLM; Textual Advertisement

## 1. Introduction

The web is flooded with human-generated texts promoting products, services, or places. These descriptive advertisements often present information in incomplete sentences, designed more to catch attention than to inform comprehensively. Identifying, extracting, structuring, and integrating information from this specialized content into an ontology. In other words, automatic ontology population can organize and, effectively, utilize this unstructured descriptive content.

Ontologies organize information in a structured manner by linking entities through semantic relationships. Ontology population involves enriching an ontology by adding concrete instances. This instantiated ontology can then be used as a knowledge graph which provides an efficient representation of knowledge, enabling machines to reason and answer complex questions.

This paper specifically addresses automatic ontology population from textual advertisements by comparing a semantic approach with an LLM-based approach.

\* Aya Sahbi. Tel.: +33231567394

E-mail address: [aya-nour-elimane.sahbi@unicaen.fr](mailto:aya-nour-elimane.sahbi@unicaen.fr)

The two main contributions of this paper are:

- We compare and present results for two ontology population approaches: KOnPoTe, a semantic method, and an LLM-based technique tested on the model Claude. Our experiments focus on two online advertisement domain ontologies: real estate and boats. To the best of our knowledge, such a comparative analysis has not yet been introduced within the semantic web community. Notably, Claude has not been widely used or tested on French data.
- The original contribution of this paper lies in its examination of short and incomplete French texts. Such a case study makes the population process challenging since the majority of related work deals with full-length English text.

This paper is organized as follows: Section 2 presents related work, Section 3 describes our case study, data and the compared approaches while Section 4 details the experiments and analyzes the results. Finally, the last section concludes and presents future work.

## 2. Related Work

Ontology population refers to the process of adding instances to an ontology, automatically or manually. Populating an ontology from text often involves Information Extraction (IE) tasks [28], such as Named Entity Recognition (NER) [39], Open Information Extraction (OIE) [26], Relation Extraction (RE) [11], and event extraction [31]. These techniques are of high interest within the Semantic Web community [2, 23, 24]. Extracting the relevant information is not the only task to be performed. The population process itself stipulates the correct association of the extracted instances, according to predefined ontological axioms, with the appropriate classes or properties. In this section, we are interested in two categories of approaches for populating ontologies from textual documents: the first category groups semantic approaches while the second reviews techniques based on Large Language Models (LLMs).

### 2.1. Ontology Population from Semantic Perspective

In this section, we examine semantic methods that focus on populating ontologies from textual documents using a domain ontology provided as an input. To name a few semantic works, [10, 13, 20, 29, 37] reviewed the different methods of ontology population. Some methods are based on Natural Language Processing (NLP) techniques and rely on Machine Learning (ML) or Deep Learning (DL) algorithms to identify entities and relationships from the text [18, 32]. These ML and DL approaches have shown remarkable performances, even when applied to different domains such as the biomedical domain or cybersecurity [8, 17]. However, they require in-depth human supervision as well as the availability of large training datasets [19, 34]. In some applications, supervised population approaches might be limited by the availability of annotated data for learning. To reduce the supervision effort required by these methods, unsupervised [8], semi-automatic [21] and hybrid [14] ontology population approaches have been introduced.

In addition to these techniques, several ontology population frameworks have been implemented, such as ArtE-quAKT [4], a web-based approach for populating ontologies in the domain of artists. Another approach by [22] focuses on extracting verbs to populate the ontology. OntoLearn [27] is a framework that extracts domain ontologies from web documents shared among members of virtual organizations.

### 2.2. Ontology Population from LLM Perspective

Large Language Models (LLMs) offer new opportunities for creating semantic representations [24, 35] or generating triples in RDF (Resource Description Framework) format [33]. These models are based on requests or so-called prompts. Prompt engineering involves designing specialized instructions to guide LLMs in generating output according to the requested specification [41].

Several LLMs have been introduced [40] (e.g., OpenAI's GPT [1], Anthropic's Claude [6], Google's Gemini (Bard) [30], etc.), where each model is trained on different datasets and developed for specific tasks. Comparative studies have been conducted to test the effectiveness of these models in different applications. The paper [3] explained that

when comparing Bard and GPT, several remarkable differences appear depending on the use case. GPT showed effectiveness in language translation tasks, product descriptions, and transcript summaries, while Bard remained more efficient in terms of information extraction, code generation and optimization. In [12], the authors investigated the ability of Claude, GPT, Bard and LLaMA in several applications and explained that in the context of named entity recognition and language comprehension, the models are almost equally effective. Several studies have analyzed the ability of these LLMs to generate populated ontologies. The paper [35] evaluated ChatGPT’s performance in extracting knowledge graphs from texts. The research work presented in [24] proposed a benchmark to evaluate the capabilities of language models to generate knowledge graphs from natural language text guided by an ontology, while [9] evaluated nine different LLMs for three ontology learning tasks: term typing, taxonomy discovery and extraction of non-taxonomic relations.

### 3. Automatic Ontology Population from Text

#### 3.1. Case Study

We illustrate our ontology population problem statement in Figure 1. Our case study focuses on the task of populating a domain ontology, given as an input, from French textual advertisements that promote an object from a given domain. The specificity of these advertisements is that they have short and incomplete sentences, exhibiting some peculiarities: sentences containing few or no verbs, object properties and data properties, few named entities and n-ary relationships<sup>1</sup>.

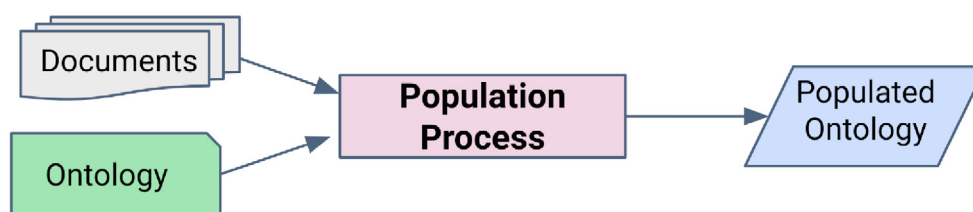


Fig. 1. Ontology population process

We previously developed KOnPoTe [5], which is an ontology population approach designed for such a context. We briefly explain this approach in Section 3.3. Although KOnPoTe has good performance in populating our advertisement domain ontology, its validation task remains time and human-resource-consuming (more details in Section 4.2). Given the related work, LLM approaches have the potential to contribute to automatically populating a domain ontology. However, it is necessary to investigate the extent to which LLMs can populate an ontology from short, incomplete text. Since we provide the full ontology as input to the LLM, we consider only LLMs that have large token size limitations. Among the state-of-the-art models that have shown interesting accuracy, Claude has the advantage of allowing a file of up to 50 MB to be attached to the prompt, which is useful in our context. For the GPT, Gemini and LLaMA models, the size of our input surpasses the token size limitation.

#### 3.2. Background and Data

Our objective is to populate a domain ontology, given as an input, with the information contained in textual documents where each document of the corpus is a textual description of an object in the domain under consideration. We study two application domains. The first one is Real estate sales: the corresponding corpus consists of house advertisements, and the ontology describes knowledge about house sales. It contains classes such as the main class

<sup>1</sup> Complex relationships involving three or more classes, for example, expressing that a real estate property is located at a distance from a given location.

Property (designating the real estate), House, Room, etc... It also contains properties, for example the object property contains (linking a real estate property to its parts, or a house to its rooms), or the data property areaInM2 (associating a house with a numerical value). The ontology also contains axioms, such as the fact that a Property can only be located in a maximum of one Municipality. Finally, Semantic Web Rule Language (SWRL) rules are defined. The second domain that we examine is boat sales. The corresponding corpus comprises boat sales advertisements, and the ontology contains classes such as the main class Boat and other parts of the boat, such as the Engine, etc... It also contains properties, for example, the object property hasEquipment, linking a boat to its equipment, or the data property enginePower associating an engine with a numerical value.

When comparing the two ontologies, we notice that the real estate one is more sophisticated than the boat one. Indeed, it contains more classes, data properties and object properties, and has a higher number of properties with domain and range class expressions. Additionally, the boat ontology (which is considered a flat ontology) has a shallower hierarchy with a depth of 2 sub-class levels and 1 sub-property level. One last distinguishing difference is that the boat ontology does not contain either n-ary relationships or SWRL rules. Regarding the textual descriptions, both corpora consist of 20 French classified advertisements per domain, extracted from advertisement websites<sup>2</sup>. While the number of documents is limited (see the explanation of the time-consuming evaluation task in Section 4.2), it is sufficient to establish a comprehensive comparison between the two approaches, as these different advertisements encompass the necessary examples to demonstrate the strengths and limitations of both approaches, thereby allowing for a comparison of the differences between them.

### 3.3. KOnPoTe

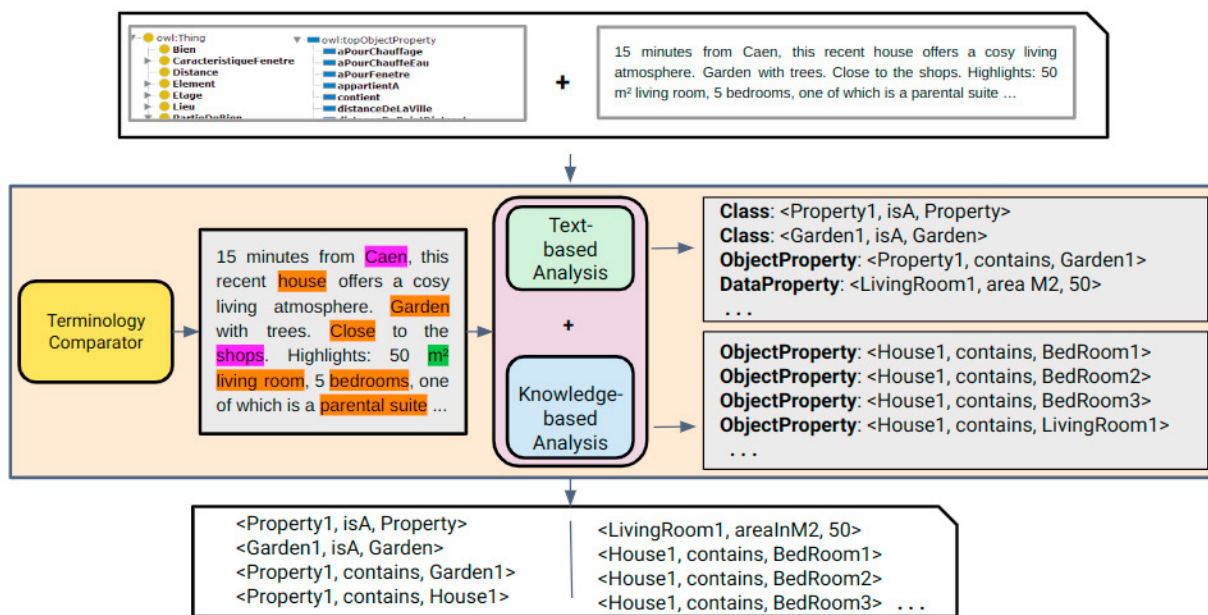


Fig. 2. Outline of the KOnPoTe approach, applied to an example document.

We developed a semantic approach that, automatically, populates a domain ontology from textual descriptions of an object. It is called "Knowledge graph/ONTology POPulation from TExts" (KOnPoTe). Figure 2 summarizes the main steps of this approach. In this section, we provide a brief explanation of KOnPoTe. More details are described in [5]. Firstly, a terminology comparator is applied to both the text and the ontology. It aims to create correspondences

<sup>2</sup> The real estate ads from <https://www.lecoindelimmo.com/>, the boat ads from <https://www.hisse-et-oh.com/>

between text mentions and ontological entities (classes, properties and individuals), taking into account the syntactic variations in keywords. For example, it detects the classes *House*, *Garden*, etc. in the document shown in Figure 2. Next, two analysis steps are implemented: a text-based analysis and a knowledge-based analysis. The text-based analysis examines the correspondences obtained and uses textual indicators to populate the ontology. First, it populates the ontology according to the types of ontological entities detected in the correspondences. For example, a match with a class instantiates that class, while a match with a property instantiates that property, using the text to find the most appropriate subject and object of assertions. Further treatments are carried out, for example, to associate individuals from successive matches within the same sentence. Only the assertions that do not generate inconsistencies are added to the ontology. In the example shown in Figure 2, this step creates, among other things, instances of the classes *House*, *Garden* and *LivingRoom*, and an assertion of the data property *areaInM2* for this room with the value 50.

The knowledge-based analysis does not exploit any textual indicators but relies on the ontology knowledge and the set of assertions already created in the text-based analysis. It aims to ensure that the graph representing object property assertions of one document is as close as possible to the shape of a star, where the center corresponds to the individual representing the subject of the document in question (for example, the real estate property described in the document, in other words, *Property1* in the example from Figure 2). In this step also, only consistent properties are added to the ontology. In the example shown in Figure 2, this step adds, among other things, the assertions of the object property *contains* that link the instance of the class *House* to the various instances of the *Room* sub-classes (*Bedroom*, *LivingRoom*, etc.).

### 3.4. LLM-based Approach

As the research area of Large Language Models (LLMs) is evolving rapidly in many ways and applications, several models and techniques have been published [25, 38, 15]. The model Claude is the most suitable regarding our data constraints: language (French) and the input size. Our method, illustrated in Figure 3, exploits the LLM Claude. Claude is part of the LLMC (Large Language Model Clara) family developed by the AI startup Anthropic. Claude has been trained on human-human dialogues as well as other diverse documents and data. It uses deep learning, particularly the transformer architecture, which has proven its effectiveness in natural language processing. Claude's conversational capabilities cover a wide range of topics and domains. This AI model stands out, in particular, for its ability to handle long queries and to take a joint file as input in addition to the prompt, which is suitable for our context. We use Claude in a zero-shot pipeline, which means that we exploit this model's pre-acquired linguistic understanding, obtained during its general training by Anthropic, without providing it with specific training for the population task. The approach takes as input a prompt containing a textual explanation of the task to be performed (in this case, populating the ontology), as well as a corpus document and the ontology. Note that this approach is equivalent to KOnPoTe in terms of input (ontology and corpus documents, processed one by one) and output (the populated ontology).

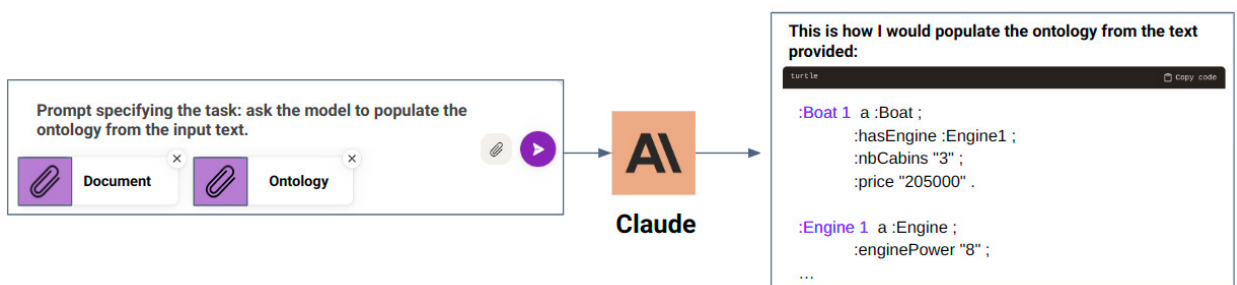


Fig. 3. Outline of the LLM-based approach, applied to an example document.

The model is interrogated directly with a zero-shot prompt, in other words, without any prior prompt tuning (few-shot, chain-of-thought prompting, etc.), to ensure a fair comparison with KOnPoTe. The approach takes as input a prompt containing a textual explanation of the task to be performed, as well as a corpus document and the ontology.

## 4. Experiments and Comparison

This section presents our experiments on the corpus of real estate and boats. KOnPoTe is tested via the executable provided in [5]. Claude is tested directly via its API<sup>3</sup>.

### 4.1. Prompt Engineering

The response generated by an LLM is highly dependent on the prompt given to the model. In this section, we detail the prompt engineering used to determine the most appropriate prompt. Several works have discussed the impact of prompt engineering on LLM performance and the variance of responses generated for the same task using different prompts. [7, 16] have shown that the response has a strong correlation with the prompt given as input. [36] has proposed a catalog of prompt engineering techniques, so-called patterns, which have been applied when interacting with ChatGPT. These techniques are proposed and tested only on the LLMs GPT and Bard. However, these patterns have not, to our knowledge, been tested on Claude, which is addressed in this paper. We have based our study on these engineering patterns and taken into account the following elements, which specify our objective and differ from related work:

- Task: populating an ontology from text and a given ontology as input.
- Language: the language used in the prompt, the ontology entities, and the text is French. We would like to precise that Claude is a multilingual model and supports French as a prompting language.
- LLM version: Claude-Instant.

#### Result 1:

```
(House1, hasHeatingSystem, gas_heating)
(House1, contains, BedRoom1)
(House1, contains, ParentalSuite1)
...
```

**Prompt 1:** Populate the ontology from the text, the ontology and text are provided as an attached file.

#### Result 3:

```
:Property1 rdf:type :Property.
:Property1 :contains :House1.
:House1 :hasHeatingSystem :gas_heating.
:House1 :contains :BedRoom1.
...
```

**Prompt 3:** Consider an ontology O representing real estate sales domain. The main class of O is the "Property" class. The text T corresponds to an ad for real estate. Files of O and T are attached. Populate the ontology O from the text T respectively according to the ontological axioms provided in O and provide the result in TTL format.

#### Result 2:

```
:House1 :hasHeatingSystem :gas_heating.
:House1 :contains :BedRoom1.
...
```

**Prompt 2:** Populate the ontology from the text, the ontology and text are provided as an attached file. Give the result in TTL format

#### Result 4:

```
:Property1 rdf:type :Property.
:Property1 :contains :House1.
:House1 :contains :BedRoom1.
:House1 :contains :BedRoom2.
...
```

**Prompt 4:** Consider an ontology O representing real estate sales domain. The main class of O is the "Property" class. The text T corresponds to an ad for real estate. Files of O and T are attached. Populate the ontology O exclusively from the text T respectively according to the ontological axioms provided in O and provide the result in TTL format.

Fig. 4. Prompt engineering applied on real estate domain

We create the prompt progressively by studying the response given by each prompt specifying to populate the ontology domain. Figure 4 illustrates the 4 prompts used and the result generated by Claude for each prompt to

<sup>3</sup> <https://www.anthropic.com/api>

populate the ontology from the document presented in Figure 2. Prompt 1 is a basic prompt that specifies the task only. It generates triples of the form (subject, predicate, object) from ontology instances. However, the format of the result of this prompt cannot be directly exploited in the validation step. To remedy this, we specify the desired format, Turtle, in Prompt 2. Analyzing the result of Prompt 2, we notice that the LLM did not respect the ontological axioms given in the input ontology, has not automatically detected the main class *Property* of the ontology, and has not generated the assertions related to this class (its parts, price, etc.). At this stage, the LLM is unable to detect this class, as it is not aware of the context. According to [36], the model lacks the context of the words used in the prompt. We have clarified this context in Prompt 3 and explicitly instructed the LLM to respect the ontological axioms. Prompt 3 populates all the classes in our ontology, but it also generates information that is not contained in the text. For example, the result of Prompt 3 mentions that the property has gas heating and an electric water heater, but this is not indicated in the advertisement text. The addition of these extra instances can be due to an LLM hallucination. To resolve this confusion, we explicitly specified in Prompt 4 to populate the ontology only with information mentioned in the input text, by adding the word "exclusively" to Prompt 3. The last prompt (Prompt 4) has been tested on both domains and seems to be the most suitable for our automatic ontology population task. This prompt is then used and evaluated according to the experimental protocol detailed hereafter.

#### 4.2. Experimental Protocol

The house sales corpus and the real estate ontology used in our tests are provided by KOnPoTe<sup>4</sup>, except that we only considered the first 20 advertisement descriptions in this corpus. The results obtained for the boat sales domain and those of the Claude-based approach on the real estate domain are available online<sup>5</sup>. They are evaluated using the same protocol as [5]. A Gold Standard (GS) is created and exploited for each domain. This GS is the initial ontology manually annotated with assertions representing the textual corpus. We compared the assertions generated by both approaches with those of the GS for both domains. We calculate for each domain precision, recall, and F1-score given by the following formulas:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

A true positive (TP) is a correct assertion, a false positive (FP) is an erroneous assertion, and a false negative (FN) is a missing assertion. Both macroscopic (mac) and microscopic (mic) values are calculated for each evaluation metric. The macroscopic score consists of averaging the metrics for each advertisement (all advertisements have the same weight), while the microscopic score takes into account the sum of all TP, FP, and FN (all assertions have the same weight). Thus, an advertisement generating many assertions has more impact on a microscopic metric, while an advertisement generating few assertions has more impact on the macroscopic calculation. We also check the consistency of the resulting ontologies using the Pellet reasoner.

We decided to launch experiments on only 20 advertisements per domain since the evaluation of both approaches is time and resource-consuming. Indeed, the creation of the GS for each domain is not the only laborious task; there is also the comparison between each advertisement assertion and the corresponding GS assertion, to determine the correct, mistaken, and missing ones. This comparison involves determining, for each instance generated by both approaches, prior equivalences with the GS instances. We also need to calculate the number of missing sameAs links in case of non-detected equivalences and ignore assertions that repeatedly represent the same property (see [5] for more explanation about this evaluation step).

<sup>4</sup> <https://zenodo.org/records/5776752>.

<sup>5</sup> <https://zenodo.org/records/11640883>.

### 4.3. Results and Discussion

Table 1 shows the results of both approaches. KOnPoTe outperforms the Claude-based approach in both domains. Indeed, KOnPoTe achieves better results on all the metrics calculated. The Claude-based approach has a relatively lower precision (77-80%), which means that a part of the assertions generated are false (around 20%). Moreover, recall is low (56-75%), which means that many assertions are missing (around 25%).

Table 1. Results of the two approaches on both domains

Domain	Approach	Precision (mac)	Recall (mac)	F1-score (mac)	Precision (mic)	Recall (mic)	F1-score (mic)
Real Estate	KOnPoTe	<b>0.94</b>	<b>0.90</b>	<b>0.92</b>	<b>0.93</b>	<b>0.87</b>	<b>0.90</b>
	Claude	0.77	0.59	0.66	0.77	0.56	0.65
Boats	KOnPoTe	<b>0.91</b>	<b>0.92</b>	<b>0.92</b>	<b>0.91</b>	<b>0.92</b>	<b>0.92</b>
	Claude	0.77	0.75	0.76	0.80	0.75	0.77

A detailed analysis of the results obtained has enabled us to deduce the strengths and limitations of both approaches. These are summarized in Table 2. KOnPoTe performs rather well, but very often generates several equivalent individuals. For example, the document in Figure 5 mentions "five bedrooms" and then "four bedrooms". KOnPoTe generates 9 instances of the class Bedroom, without taking into account the fact that the four bedrooms are included in the five bedrooms mentioned in the previous sentence. Conversely, Claude does not generate this kind of redundancy of individuals but essentially generates some inconsistent assertions in terms of reasoning. Indeed, the ontology generated by Claude is inconsistent for 90% of the real estate advertisements and 60% of the boat advertisements. This is because the domain and range axioms are not always respected by this approach. On the other hand, this kind of error is not possible with KOnPoTe, as each step of the approach ensures that candidate assertions do not generate inconsistencies, and does not add them if so. When it comes to assertions of n-ary properties, neither approach works perfectly, but KOnPoTe is doing well, while Claude has more difficulty, particularly in linking individuals together. The LLM has the advantage of detecting abbreviations in the text, while KOnPoTe is not capable of detecting them unless they are mentioned in the input ontology. Finally, both approaches fail in the most complicated cases, which require a strong understanding of the text.

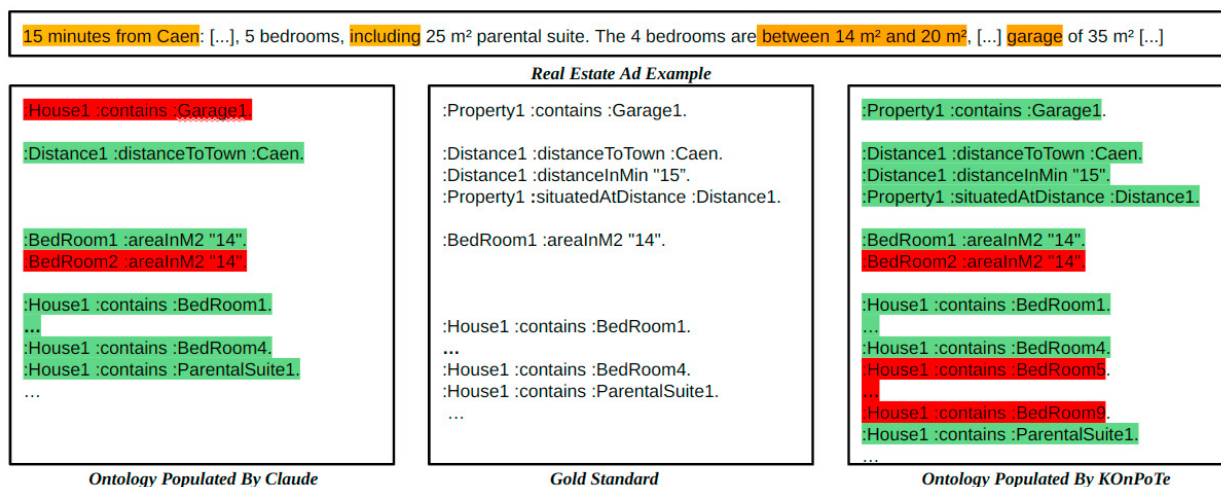


Fig. 5. Example of real estate ontology populated by Claude and KOnPoTe from a house sale ad and the corresponding Gold standard

Figure 5 shows an example of a real estate advertisement, the assertions generated by both approaches, and the GS. When analyzing the examples, we notice that Claude has generated some mistaken assertions (object properties with wrong range and domain) compared to the GS and KOnPoTe. Indeed, the real estate ontology provided as input

specifies that only instances of `Property` can contain instances of `PropertyPart`. The class `Garage` is a subclass of `PropertyPart`, so `Garage1` described in Figure 5 can only be contained by a `Property`. However, we can see that the individual `Garage1` is contained by an instance of `House`, and not an instance of `Property`, which makes the ontology inconsistent according to the Pellet reasoner. On the contrary, KOnPoTe has created a consistent and correct assertion.

An example of n-ary relationship is illustrated in Figure 5. The example shows that the real estate is located 15 minutes from Caen. We can see that the population is correctly done by KOnPoTe: `Property1` is located at a distance `distance1`, and `distance1` is associated with the notion of 15 minutes and with the city of Caen. For Claude, the `distance1` instance created does indeed concern the city of Caen, but it is neither associated with `Property1` nor with the notion of 15 minutes.

On the other hand, KOnPoTe tends to generate redundant individuals. For example, it generates 9 bedrooms instances and one parental suite instance, whereas Claude generates, rightly, only 4 bedrooms instances and one parental suite instance.

Table 2. Comparison of the strength and limitations of Claude and KOnPoTe

Approach	Hilights	Limitations
KOnPoTe	<ul style="list-style-type: none"> <li>- Consistency of results</li> <li>- Explainability of results</li> <li>- Detection of n-ary relationships</li> </ul>	<ul style="list-style-type: none"> <li>- Redundancy of generated individuals</li> <li>- Confusion in complicated cases</li> </ul>
Claude	<ul style="list-style-type: none"> <li>- No redundancy of individuals</li> <li>- Detection of abbreviations</li> </ul>	<ul style="list-style-type: none"> <li>- Inconsistency of results</li> <li>- Inexplainability of results</li> <li>- Confusion in complicated cases</li> </ul>

Finally, Table 1 shows that KOnPoTe's results are better for real estate than boats, whereas it is the opposite for Claude. We deduce that the KOnPoTe approach performs better when the ontology to be populated is more sophisticated, whereas Claude populates better simpler ontologies.

## 5. Conclusions and Future Work

This paper presents a comparative study of two different approaches of automatic ontology population: a semantic one and an LLM-based one. This study shows that each approach has its qualities and limitations. Firstly, the semantic approach is more efficient, exploiting the reasoning capacity of ontologies. However, it does have its limitations, notably in terms of individual redundancy. The LLM approach is less efficient when dealing with consistency. However, it does not tend to create redundancies for individuals. In addition, KOnPoTe is adapted to the context of French descriptive texts (the use case tested). However, another context might require the implementation of another approach. As Claude's use is simplified and straightforward via an available API, it remains more immediate than the design and implementation of a semantic approach adapted to the context of the documents exploited.

An interesting perspective would be to propose a hybrid approach, adapted to various contexts, based on the results of the LLM approach and seeking to correct inconsistent assertions generated thanks to ontology knowledge. Such an approach could, then, benefit from a module inspired by KOnPoTe's knowledge-based analysis to add consistent missing assertions.

## References

- [1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .
- [2] Adnan, Kiran, R., 2019. An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data* .
- [3] Ahmed, I., Kajol, M., Hasan, U., Datta, P.P., Roy, A., Reza, M.R., 2023. ChatGPT vs. Bard: A Comparative Study. *UMBC* .
- [4] Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R., 2003. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems* 18, 14–21.
- [5] Alec, C., 2023. Ontology population from french classified ads, in: *International Conference on Conceptual Structures*, Springer. pp. 155–170.
- [6] Anthropic, A., 2024. Introducing the next generation of claude.

- [7] Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M., Glassman, E., 2023. Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. preprint arXiv:2309.09128 .
- [8] Ayadi, A., Samet, A., de Beuvron, F.d.B., Zanni-Merk, C., 2019. Ontology population with deep learning-based NLP: a case study on the biomolecular network ontology. *Procedia Computer Science* 159, 572–581.
- [9] Babaei Giglou, H., D'Souza, J., Auer, S., 2023. Llm4ol: Large language models for ontology learning, in: ISWC, Springer. pp. 408–427.
- [10] Biemann, C., 2005. Ontology learning from text: A survey of methods. *Journal for Language Technology and Computational Linguistics* .
- [11] Boer, V.D., Someren, M., Wielinga, B., 2006. Relation instantiation for ontology population using the web , 202–213.
- [12] Borji, A., Mohammadian, M., 2023. Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard .
- [13] Buitelaar, P., Cimiano, P., Magnini, B., 2005. Ontology learning from text: methods, evaluation and applications. volume 123. IOS press.
- [14] Castano, S., Peraldi, I.S.E., Ferrara, A., Karkaletsis, V., Kaya, A., Möller, R., Montanelli, S., Petasis, G., Wessel, M., 2009. Multimedia interpretation for dynamic ontology evolution. *Journal of Logic and Computation* 19, 859–897.
- [15] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al., 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* .
- [16] Gao, A., 2023. Prompt engineering for large language models. Available at SSRN 4504303 .
- [17] Gasmi, H., Laval, J., Bouras, A., 2019. Cold-start cybersecurity ontology population using information extraction with Istm, in: 2019 International Conference on Cyber Security for Emerging Technologies (CSET), IEEE. pp. 1–6.
- [18] Imsombut, A., Sirikayon, C., 2016. An alternative technique for populating thai tourism ontology from texts based on machine learning, in: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), IEEE. pp. 1–4.
- [19] Jayawardana, V., Lakmal, D., de Silva, N., Perera, A.S., Sugathadasa, K., Ayesha, B., Perera, M., 2017. Semi-supervised instance population of an ontology using word vector embedding, in: Int. Conference on Advances in ICT for Emerging Regions (ICTer 17), IEEE. pp. 1–7.
- [20] Lubani, M., Noah, S.A.M., Mahmud, R., 2019. Ontology population: Approaches and design aspects. *Journal of Information Science* 45.
- [21] Maedche, A., Staab, S., 2001. Ontology learning for the semantic web. *IEEE Intelligent systems* 16, 72–79.
- [22] Makki, J., Alquier, A.M., Prince, V., 2009. Ontology population via nlp techniques in risk management. *Int. Journal of Humanities and Social Science (IJHSS)* 3, 212–217.
- [23] Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I., 2020. Information extraction meets the semantic web: a survey. *Semantic Web* 11.
- [24] Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K., 2023. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text, in: International Semantic Web Conference, Springer. pp. 247–265.
- [25] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2024. Large language models: A survey. arXiv:2402.06196 .
- [26] Muhammad, I., Kearney, A., Gamble, C., Coenen, F., Williamson, P., 2020. Open information extraction for knowledge graph construction, in: Database and Expert Systems Applications: DEXA 2020 International Workshops BIODDD, IWCFS and MLKgraphs, Bratislava, Slovakia, September 14–17, 2020, Proceedings 31, Springer.
- [27] Navigli, R., Velardi, P., 2004. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics* 30.
- [28] Nédellec, C., Nazarenko, A., 2006. Ontologies and information extraction. arXiv preprint cs/0609137 .
- [29] Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E., 2011. Ontology population and enrichment: State of the art. *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap* , 134–166.
- [30] Pichai, S., Hassabis, D., 2023. Introducing gemini: our largest and most capable ai model .
- [31] Sahnoun, S., Elloumi, S., Ben Yahia, S., 2020. Event detection based on open information extraction and ontology. *Journal of Information and Telecommunication* 4, 383–403.
- [32] Sambandam, P., Yuvaraj, D., Padmakumari, P., Swaminathan, S., 2023. Spiking equilibrium convolutional neural network for spatial urban ontology. *Neural Processing Letters* , 1–20.
- [33] Souza, R.R., Pinheiro, T., Oliveira, J., dos Reis, J.C., 2023. Knowledge graphs extracted from medical appointment transcriptions: Results generating triples relying on llms , 129–139.
- [34] Suchanek, F., Ifrim, G., Weikum, G., 2006. Leila: Learning to extract information by linguistic analysis, in: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 18–25.
- [35] Trajanoska, M., Stojanov, R., Trajanov, D., 2023. Enhancing knowledge graph construction using large language models. arXiv preprint arXiv:2305.04676 .
- [36] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C., 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. preprint arXiv:2302.11382 .
- [37] Wong, W., Liu, W., Bennamoun, M., 2012. Ontology learning from text: A look back and into the future. *ACM computing surveys (CSUR)* .
- [38] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., Hu, X., 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* .
- [39] Zhang, Z., Ciravegna, F., 2011. Named entity recognition for ontology population using background knowledge from wikipedia, in: Ontology learning and knowledge discovery using the web: Challenges and recent advances. IGI Global, pp. 79–104.
- [40] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., rong Wen, J., 2023. A survey of large language models. ArXiv abs/2303.18223. doi:10.48550/arXiv.2303.18223.
- [41] Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J., 2022. Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910 .