Enhancing causal graphs with domain knowledge: matching ontology concepts between ontologies and raw text data

Jernej Stegnar
Jožef Stefan Institute
Ljubljana, Slovenia
jernej.stegnar@gmail.com

Jože M. Rožanec
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Gregor Leban
Event Registry d.o.o.
Ljubljana, Slovenia
gregor@eventregistry.org

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

When building a causal graph from textual sources, such as media reports, a key task is to provide an accurate semantic understanding of the causal variables encoded as nodes and to link them to existing ontologies with at least two purposes: (i) expand the knowledge with the domain knowledge captured in such ontologies and (ii) provide accurate and different levels of abstraction of the extracted causal variables. This article describes how we used OntoGPT, a tool for matching raw text to ontology concepts initially designed for the medical domain, to match concepts from media events to relevant ontologies. We build upon our previous work on extracting causal variables and enrich the extraction pipeline by matching causal variables to concepts from specific domain ontologies. In particular, we describe our work regarding the GEO ontology. Future work will focus on expanding OntoGPT's capabilities by utilizing a wider selection of ontologies. Addressing its limitations, such as dealing with multiple instances of the same class, will also be crucial for improving its utility. These improvements will allow the tool to better support strategic foresight applications by providing more detailed insights across a multitude of different sectors, further enriching causal graphs and facilitating more accurate predictive modeling.

KEYWORDS

strategic foresight, ontology matching, artificial intelligence

1 INTRODUCTION

Strategic foresight is a discipline concerned with anticipating future trends, uncertainties, and disruptions to inform decision-making and enable the creation of resilient, long-term strategies. As such, it is valuable to governments, organizations, and enterprises, who can use it to remain competitive and adaptable in a rapidly changing world [4].

The pace of technological advancement, shifting geopolitical landscapes, environmental crises, and unpredictable market trends make it essential to react quickly to change. Traditionally, foresight has been based on trend analysis, expert opinion, and qualitative insights. Such approaches lack the agility required to scan real-world events in near-real time and produce strategic foresight outcomes at such a pace. Nevertheless, this would be possible with the use of artificial intelligence.

AI enhances strategic foresight by automating the analysis of data and detecting patterns that may go unnoticed by human experts [1]. Machine learning algorithms can continuously monitor emerging trends, geopolitical shifts, and market fluctuations in near-real time, offering dynamic insights into potential future scenarios. Natural language processing (NLP) enables AI to sift through massive amounts of text, extracting relevant information from reports, news, and social media, thus accelerating the forecasting process. By integrating AI into strategic foresight, organizations can adapt more swiftly and make more informed, data-driven decisions in the face of uncertainty.

Ontologies provide structured knowledge informing the relationships between concepts within a specific domain. Furthermore, they describe those concepts through properties and can link such classes to specific instances observed in the real world. As such, they are of key importance when building a causality graph, given they can augment our understanding of the causal relationships between variables with a better understanding of the context and the variable implications [3]. For example, if the causal relationship reports about the ceasing of an armed conflict, knowing whether a causal variable relates to a country, the location of that country, the neighboring countries, and international organizations it is involved in would help to understand the magnitude of that event and contextualize other likely outcomes (refugee repatriation, impacts on investments, and others).

In the scope of the graph massive project, ontology matching is being used to link the extracted causal relationships from text to concepts inside the ontologies, allowing for a more detailed understanding of the concepts that appear in causal relationships and their interconnectivity.

2 ENRICHING CAUSAL GRAPHS WITH DOMAIN KNOWLEDGE

We consider ontologies a framework (an organized and structured system for representing knowledge) used to represent knowledge within a specific domain by defining the relationships between concepts. They consist of classes (concepts), properties (attributes), and relationships that connect different concepts. This structure provides a standardized way to organize and interpret data, ensuring consistent understanding across systems. For example, in a medical ontology, concepts like "disease" might be linked to "symptoms," "treatments," and "causes," each with its own defined properties. By formalizing these relationships, ontologies allow AI systems to better interpret and reason about

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.25

complex information, leading to more accurate data processing and decision-making.

Ontologies enhance causality graphs by providing domain-specific knowledge that improves the accuracy and depth of relationships represented. When extracting causal relationships from large datasets, such as media reports, the data can often be ambiguous or incomplete. Ontologies address this by offering structured knowledge that defines concepts and their relationships within a specific domain, linking extracted causal relationships to well-defined entities in the ontology. This enriches the causality graph, uncovering implicit connections and non-obvious relationships that may otherwise be missed. In strategic foresight, for example, ontology-based enrichment helps capture a broader range of potential future scenarios by incorporating knowledge beyond the immediate dataset. This leads to more reliable predictions, especially when the training data is limited or domain-specific. Ultimately, ontologies are expected to enable the system to generalize better, predict outcomes with higher accuracy, and improve the overall reliability of causality graphs.

The causality graph pipeline in the Graph Massivizer strategic foresight project is designed to automate the extraction, organization, and analysis of causal relationships from large datasets, particularly news articles. The Figure 1 showcases the structure of our causality graph's data pipeline. The process begins with extracting these relationships from news articles, which are then organized into a causality graph that maps the interactions between various factors and events. The goal is to develop link prediction models that estimate the likelihood of future events based on observed patterns. For instance, one use case involves predicting oil price trends by analyzing factors that influence pricing.

Ontology matching is then integrated into the pipeline to link extracted causal relationships with concepts from structured ontologies. This enrichment adds layers of context and enables the discovery of connections that may not be evident from raw data alone. By incorporating ontologies, the pipeline transcends the limitations of its training data, identifying causal relationships that may be implied by broader knowledge contained in the ontologies. This not only enhances the accuracy of the graph but also allows it to capture more complex and non-direct relationships, improving its predictive capabilities.

As shown in Fig. 1B, the process of ontology linking in our pipeline consisted of creating ontology matching templates, then linking the concepts in text to ontologies, using the information to add additional data to existing causalities, all with the purpose of finding extra implicit connections based on the information provided by the ontologies.

The main problem that needed solving for that purpose was, how to link ontologies to raw text data. In our case that was done using OntoGPT [2], which is a tool used for ontology linking. Another key challenge is inter-ontology matching, which involves linking multiple ontologies through shared concepts. This process expands the knowledge framework, making it even more valuable for our purposes. The challenge of inter-ontology matching hasn't been addressed yet and remains a matter of future work.

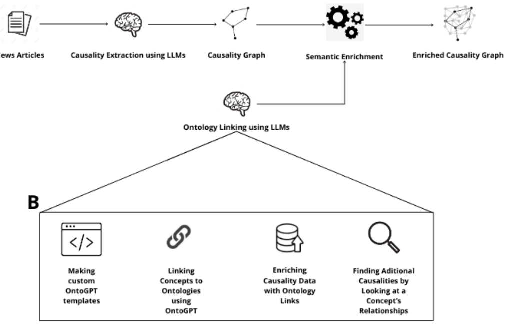

News Articles
Causality Extraction using LLMs
Causality Graph
Semantic Enrichment
Enriched Causality Graph
Ontology Linking using LLMs
B
Making custom OntoGPT templates
Linking Concepts to Ontologies using OntoGPT
Enriching Causality Data with Ontology Links
Finding Aditional Causalities by Looking at a Concept's Relationships

Figure 1: The figure showcases our pipeline for building a causality graph. The sub-figure B showcases how the process of ontology linking was executed as a part of our pipeline

3 ONTOGPT: A BRIEF OVERVIEW

OntoGPT is an advanced tool that integrates large language models (LLMs) with ontologies to improve knowledge extraction and organization across various domains. Ontologies provide a consistent and accurate representation of complex information by defining structured relationships between concepts.

The primary purpose of OntoGPT is to enhance AI systems' understanding, processing, and categorization of data by linking extracted information to predefined concepts and relationships within an ontology. This structured approach ensures greater accuracy and reliability compared to traditional AI systems that rely on unstructured data.

OntoGPT works by connecting data from sources such as text or reports to specific concepts in an ontology, allowing for more informed and contextually accurate connections. For example, in healthcare, OntoGPT can link symptoms from patient records to diseases and treatments outlined in medical ontologies, helping to suggest possible diagnoses or treatment plans.

By combining the language-processing capabilities of LLMs with the structured knowledge available in ontologies, OntoGPT enables AI systems to go beyond keyword matching and consider the relationships between terms. This leads to more intelligent data interpretation and improved decision-making.

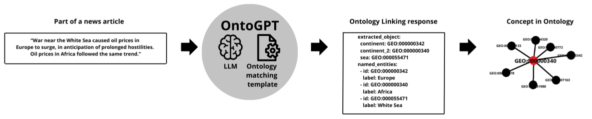
OntoGPT is widely used in fields where structured knowledge is critical for high accuracy, such as healthcare, biology, and pharmaceutical research. In medical research, for instance, OntoGPT links clinical trial data, medical records, and scientific literature to medical ontologies, supporting better analysis and decision-making.

The key advantage of OntoGPT lies in its ability to ground AI outputs in domain-specific, structured knowledge, reducing the likelihood of errors and improving the relevance of insights. This grounding ensures that AI responses are not just based on patterns but also on well-defined concepts and their relationships.

In summary, OntoGPT bridges the gap between the raw data-processing power of LLMs and the structured knowledge in ontologies. By leveraging both, it provides a more accurate and reliable approach to extracting and linking data across various domains, particularly when working with large, complex datasets.

3.1 OntoGPT's role

At a lower level, OntoGPT operates using YAML templates that define how data should be extracted from text and linked to ontological concepts. These templates serve as blueprints, specifying which types of entities, relationships, and properties to look for in the input text. The templates guide the large language model by mapping textual data to predefined concepts and relationships



"War near the White Sea caused oil prices in Europe to surge, in anticipation of prolonged hostilities. Oil prices in Africa followed the same trend."

```
extracted_object:
  continent: GEO:000000342
  continent_2: GEO:000000340
  sea: GEO:000055471
named_entities:
  - id: GEO:000000342
    label: Europe
  - id: GEO:000000340
    label: Africa
  - id: GEO:000055471
    label: White Sea
```

Figure 2: A Showcase of the function of OntoGPT

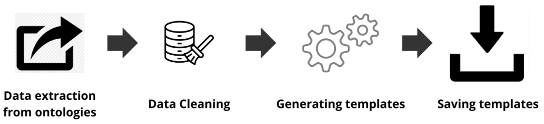


Figure 3: The Process of Templates Generation

from the ontology, ensuring that the extracted information is both relevant and structured. The figure 2 shows the process of ontology linking for an example of a simple sentence. Each YAML template contains detailed instructions on how to identify key terms, their corresponding ontology classes, and the relationships between them. This allows OntoGPT to recognize when a piece of text, such as a sentence from a media article, contains a concept that aligns with an entity or event in the ontology. Once identified, the tool links the extracted data to these ontology entries, enabling richer and more meaningful connections in the data, as it is now grounded in an established knowledge framework.

The approach described in this article uses an ontology file as input to create such templates for data extraction and linking. This enables for a broader range of ontology linking, as the templates can be created on demand.

4 TEMPLATES AND PYTHON CODE GENERATION

The approach works by using the information defined inside the ontology, to generate the YAML templates. The Figure 3 showcases the process of how this is done.

First the class information, for each class inside the ontology, is extracted. This is done by using the "owlready2" python library to parse the ontology into an object, and then extract the relevant information from the new object.

Every class inside the ontology is used to create a corresponding template class, which is optimal, as it covers all parts of the ontology that could potentially be linked. A small portion of the data extraction process is ontology-specific and was custom-tailored to the individual ontology, as some information (like class descriptions) is saved in different parts.

Secondly the data extracted from the ontology is processed and used to create custom YAML templates. This is done by simply using the extracted information to fill in a "general template" we used for generation. Specifically the class names and descriptions are used, to do so. This gives OntoGPT the names of the classes inside the ontology, that we are trying to link the text data to, and their descriptions, which assists OntoGPT in more accurately identifying these classes inside the text. The YAML file also contains the information of "annotators" which tells OntoGPT, which ontology to ground the responses to. The generated YAML templates are saved into a separate file after generation, which makes them ready for use.

The python code that is used by OntoGPT in the process of ontology linking, is similarly generated by using the extracted information to fill in the "general template" and is then saved to a separate file.

5 LIMITATIONS

5.1 Multiple Same-Class Concepts

OntoGPT has problems trying to link two or more concepts to a place in the ontology if the concepts are of the same class. This happens because both concepts suit the description and similar criteria that OntoGPT extracts the information based on. This causes OntoGPT to merge both concepts into a single string and then try to locate the said string inside the ontology, which fails because there is no individual inside the ontology class with such a name. An example of such a response is shown in Listing 1:

Listing 1: Example of a bad response

```
extracted_object:
  continent: AUTO:Europe%2C%20Africa
named_entities:
  - id: AUTO:Europe%2C%20Africa
    label: Europe, Africa
```

If OntoGPT managed to locate the concept inside the text in the ontology, it returns its id (an example of this is "sea: GEO:000055471" and "id: GEO:000055471 : White Sea") If the concept suits the class criteria, but couldn't be located inside the ontology, it returns it as a "AUTO" detection. For the purpose of ontology linking this is not optimal as it does not give us access to the additional information that is stored inside the ontology's individual information. The ontology's individual information is a set of predefined relationships and properties, that an individual concept has. For example, if the individual "Africa" is defined inside the ontology, the individual's data would include its size, countries on the continent, population, and climates, among others. This information gives us reliable information about a certain concept, allowing for more contextual understanding.

To solve this problem, the approach of creating "buffer" classes was taken, where a certain class from ontology would be used to generate three classes describing the different occurrences of the ontology class and a description that would provide sufficient context to OntoGPT to separate the same class concepts into different entities. The corrected response is showcased in Listing 2:

Listing 2: Example of a corrected response

```
extracted_object:
  continent: GEO:000000340
  continent_2: GEO:000000342
named_entities:
  - id: GEO:000000340
    label: Africa
  - id: GEO:000000342
    label: Europe
```

While this approach deals with a high percentage of this type problem, it does not cover the cases where more than three same-class concepts are inside the piece of text being analyzed.

6 CONCLUSIONS

Using OntoGPT in the Graph Massivizer strategic foresight project will prove valuable for enriching causal graphs with linked ontology data, aiming to improve predictive accuracy in predicting future events. Despite OntoGPT's initial focus on medical data, some custom adaptations were successfully implemented to suit a portion of different domains. However, limitations persist in distinguishing between multiple instances of the same concept class. These challenges highlight the need for further development to enhance the tool's versatility across a broader array of applications and ontologies.

ACKNOWLEDGMENTS

The Slovenian Research Agency supported this work. This research was developed as part of the Graph-Massivizer project funded under the Horizon Europe research and innovation program of the European Union under grant agreement 101093202.

REFERENCES

[1] Patrick Brandtner and Marius Mates. 2021. Artificial intelligence in strategic foresight–Current practices and future application potentials: current practices and future application potentials. In *Proceedings of the 2021 12th International Conference on E-business, Management and Economics*. 75–81.

[2] J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, et al. 2024. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics* 40, 3 (2024), btae104.

[3] Fatma Özcan, Chuan Lei, Abdul Quamar, and Vasilis Efthymiou. 2021. Semantic enrichment of data for AI applications. In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning*. 1–7.

[4] David Sarpong and Nicholas O'Regan. 2014. The Organizing Dimensions of Strategic Foresight in High-Velocity Environments. *Strategic Change* 23, 3-4 (2014), 125–132.