Enhancing Automatic Ontology Engineering: An Exploration of Integrating Topic Extraction Algorithms and Large Language Models

Ömer Ülgen 2690752

Vrije Universiteit Amsterdam

1 Abstract

This thesis explores the challenges and opportunities inherent in the application of ontologies in the domain of clinical trials, particularly concerning Gastrointestinal (GI) cancers. The research is driven by the primary objective of developing an automated method for creating a GI cancer-specific ontology to improve knowledge accessibility and usability in this domain. The study's focus is on addressing issues related to data heterogeneity and interoperability, which have hindered efficient data management and integration in the context of clinical trials. This research utilizes Latent Dirichlet Allocation (LDA) [1] and OPEN AI's API [2] to automate ontology generation, aiming to streamline knowledge representation, integration, and interoperability. The effectiveness of these techniques is evaluated against a manually crafted ontology. The study seeks to answer the research question: How do ontology engineering approaches, incorporating Large Language Models , compare in quality to manually crafted ontologies? By addressing this question, the study aims to shed light on the potential benefits and limitations of automated ontology generation techniques and pinpoint opportunities to improve the quality, efficiency, and utility of ontologies for GI cancers and beyond. The developed ontologies' coverage is evaluated using manual mappings of major classes.

In the course of this research, it was discovered that the combined utilization of Topic Extraction Algorithms and Large Language Models (LLMs) yielded satisfactory results in ontology generation. Notably, the best model (TF-IDF) correctly mapped 10 out of 16 major classes, and the BOW model fully covered 6 out of the 16 major classes from the golden ontology . This indicates a potential for automated ontology generation techniques to facilitate knowledge representation and semantic interoperability effectively. However, it was observed that the nature of the applied algorithms and experiments did not fully reflect the hierarchical structure typically found in manually constructed ontologies. Instead, they tended to represent the discovered classes within a single broad category.

2 Introduction

In recent years, the development and implementation of ontologies has become essential in various domains, including healthcare, to address the growing need for structured knowledge representation and semantic interoperability. Clinical trials play a pivotal role in the healthcare industry, as they constitute the cornerstone of the drug development process, providing the evidence needed for regulatory approval and subsequent prescription to patients. The usage of these ontologies within the clinical trial domain ranges from: matching patient records with external clinical trials [3], integration of clinical trials for management applications [4, 5], automatic information extraction from past trials [6], and many other implementations. Hence, the general consensus supports the fact that employing ontologies in the field of clinical trials is fundamental to ensuring an efficient and effective clinical trial system [7].

Gastrointestinal cancers represent a significant global health burden, and advancements in this field require a semantic framework that can facilitate efficient knowledge discovery and integration. However, ontologies associated with clinical trials face several challenges that hinder the efficient management and integration of this set of data within a unified knowledge representation that covers a global understanding of different works. These problems range from data heterogeneity to interoperability.

Data heterogeneity arises from the diverse nature of clinical trial data, ranging from structured information such as patient demographics and clinical outcomes to unstructured data such as free-text narratives and imaging studies. This heterogeneity makes it difficult to integrate and analyze data across multiple trials and sources, leading to potential inconsistencies and inefficiencies in the drug development process.

Interoperability is another challenge in the clinical trial domain, as data standards and terminologies often vary across organizations, making it harder and less convenient to share and compare information.

In light of these challenges, there is a need for a comprehensive ontology framework that can streamline knowledge representation, integration, and interoperability for Gastrointestinal (GI) cancers.

Therefore, the primary objective of this research is to develop an automatic method for the creation of a GI cancer-specific ontology that will enable researchers, clinicians, and decision-makers to seamlessly access, understand, and utilize the wealth of information available in this domain.

To this end, this thesis intends to employ Latent Dirichlet Allocation (LDA) and OPEN AI's API as mechanisms for automating ontology generation. The produced ontology will then be compared against a manually crafted ontology derived from identical data sources, in an effort to evaluate the effectiveness of these automatic ontology engineering techniques.

The research question guiding this study is: How do ontology engineering approaches, which incorporate the usage of Large Language Models such as: OPEN API, compare in terms of quality to ontologies that are crafted manually? By addressing this question, we aim to provide valuable insights into the potential benefits and limitations of automated ontology generation techniques and to identify opportunities for enhancing the quality, efficiency, and utility of ontologies in the domain of GI cancers and beyond.

To evaluate the quality of the resulting ontology, we compared the automatically generated ontology to a manually modeled one. The comparison is done at a high level to measure the coverage.
Firstly, we elaborate on the construction of the 'golden ontology', crafted manually to serve as a benchmark for comparison. Following this, we delve into the design choices made during the research process. Here, we shed light on the considerations that guided our approach and justified the selection of specific methods over others.
The subsequent section of the thesis explicates the methodology, emphasizing the use of Latent Dirichlet Allocation (LDA) and OpenAI API's Large Language Models (LLMs) for automatic ontology generation. It offers detailed insights into the algorithms applied, along with the rationales for their usage.
In the final sections of the thesis, we present the results and findings. We assess the outcome of the automatic ontology generation process, detailing the extent to which the generated ontologies mirrored the manually crafted 'golden ontology'.

3 Related Work

Shankar et al.[4] propose that to efficiently build and construct a sound ontology within the clinical trial data domain, the architecture of the knowledge base should 'support three types of methods': a knowledge acquisition technique, an ontology database mapping formula, and a concept-driven strategy . We will make use of these rules whilst manually building our golden GI ontology

Another approach proposed by Lin et al. [8] is a contextual-based ontology modeling method within a database domain. The method proposed by the authors transforms schemas into a natural language format with the help of a set of syntax that reshapes knowledge-based semantics; this approach could be useful for semantic translation purposes.

Ontology-based automatic categorization is also a key part of this thesis. Liu et al. [9] have expanded on the idea that the use of different terminologies and various terms affects the efficacy of autonomous categorization techniques for clinical trial-based ontologies.

D. Lonsdale et al. [10] propose a process to generate domain ontologies from text documents. Their methodology requires the use of three types of knowledge sources: a general and well defined ontology for the domain (Golden), a dictionary or any external resource to discover lexical and structural relationships between terms and a consistent set of training text documents. With these elements they are able to automate the creation of a new sub-ontology of the more general ontology, this paper's methodology and approach suits our research purpose and will be used to some extent for automatic class creation .

Other approaches that will not be used fully but partially include Latifur Khan et al. [11] who demonstrate a method to automatically construct an ontology from a set of text documents . Their overall mechanism is as follows: 1) terms are extracted from documents with text mining techniques; 2) documents are grouped hierarchically and then: 3) assign concepts to a tree graph starting from leaf nodes . This experience introduces a new bottom-up approach for ontology generation which will only partially be applied because of the complexities encountered in the later stages of the method mentioned

Finally, Bedini et al. [12] discuss different methods for automatic ontology engineering. This paper lists and compares all shortcomings and benefits of various techniques for each step of the automation process of an ontology, ranging from knowledge extraction to data generation.

4 Methodology and Approach

4.1 Golden ontology engineering method and design choices

Data Gathering: In the pursuit of creating a comprehensive and relevant golden ontology, our primary data source was the National Institute of Health [1]. The criteria for selecting studies from this resource were stringently set to ensure high data integrity and relevance. Specifically, only studies that were concluded and had documented results by 2022 were considered. This approach was employed to provide a stable and definitive dataset, devoid of the potential changes and uncertainties inherent to ongoing studies.

In view of the fact that automated ontology generation methods have to be fed a significant enough dataset as stated by Nguyen et al. [12] , the used dataset comprises 4763 clinical trial documents that range from study title to study results. This expansive dataset was chosen for the automated method to fully exploit the capabilities and the efficacy of automatic ontology engineering techniques, which can effectively process and extract meaningful insights from large volumes of data. Furthermore, the broad dataset facilitates a more comprehensive representation of the knowledge domain, potentially revealing relationships and structures that may not be immediately evident in a smaller dataset.

Data Preprocessing: Given the complexity and extensive nature of the raw data, initial steps involved sorting and cleaning the data set to retain only the most relevant information. We chose to streamline our focus on the summaries of all clinical trials, as they encapsulate the core information about each trial, providing concise yet comprehensive insights into their unique contexts and outcomes.

Further refinement of the data involved the removal of *stop words*. These are commonly occurring words like: *and*, *the*, *is*, *that*, while integral to human

[1] (https://clinicaltrials.gov/)

language, often contribute little to machine understanding of text and can create noise in the modeling process. By eliminating these *stop words*, we enhanced the models' ability to focus on more meaningful and informative words within each summary.
An additional key preprocessing step was the application of *lemmatization.* This process involves reducing words to their base or root form (lemma), which helps in consolidating different grammatical variants of a word into a single item. For instance, the words: *running*, *runs*, *ran* all reduce to the lemma 'run.' *lemmatization* assists in standardizing words and reducing redundancy in the dataset, thereby improving the efficiency and accuracy of the subsequent topic modeling.

Golden Ontology (GI cancer): The construction of the golden ontology was guided by a hybrid approach, blending elements of both top-down and bottom-up methodologies.At the outset, the recently published Clinical Measurement Ontology (CMO) [2] was utilized as a starting point in our top-down approach. The CMO ontology, with its extensive collection of classes and relations related to clinical measurements, provided a solid foundational structure for our ontology. However, rather than indiscriminately incorporating the entire CMO, we tailored the ontology to our specific needs by selecting only the relevant classes. The determination of relevance was informed by an initial examination of the Gastrointestinal (GI) cancer clinical trial dataset, ensuring the chosen classes reflected the prevalent concepts and themes within these studies. This approach can be viewed as a bottom-up methodology, where the design was influenced by the specific needs derived from our dataset.

Upon sorting and reshaping these pertinent classes, we proceeded to enrich the ontology by considering 50 randomly selected GI clinical trials . The goal of this step was to capture any clinical measurements not initially included in the CMO-derived ontology specific to the GI cancer domain.

[2] (https://bioportal.bioontology.org/ontologies/CMO/?p=summary)

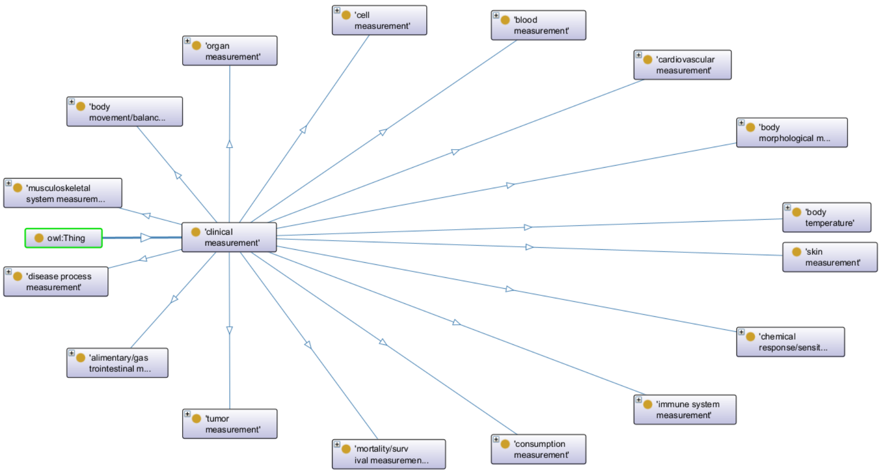

'cell measurement'
'blood measurement'
'organ measurement'
'cardiovascular measurement'
'body movement/balanc...
'body morphological m...
'musculoskeletal system measurem...
'body temperature'
owl:Thing
'clinical measurement'
'skin measurement'
'disease process measurement'
'chemical response/sensit...
'alimentary/gastrointestinal m...
'immune system measurement'
'tumor measurement'
'mortality/survival measuremen...
'consumption measurement'

Fig. 1. High level overview of golden ontology

Design Choices (GI cancer): The design choices underpinning the golden ontology aimed at achieving simplicity, structure, and compatibility with the automated ontology construction. A foundational design choice was to maintain the hierarchical structure inherited from the Clinical Measurement Ontology (CMO). All classes within the golden ontology were subsumed under a single overarching class named *Clinical Measurement* . This choice was inspired by the advantageous structure of the CMO ontology, providing a high-level organization that facilitates clear and concise knowledge representation. The hierarchical arrangement not only creates an easy-to-navigate structure but also maintains a holistic perspective of the ontology, linking every subclass to the overarching concept of clinical measurements.Another key design decision involved the naming convention for the classes. Each class within the golden ontology commences with the term Measurement, reflecting the nature of the data derived from the clinical trial documents. This choice was made to enhance the comparability between the manually created golden ontology and the automated ontology generated. By prefacing all class names with *Measurement*, we create a semantic overlap with the output of the automated ontology construction, which yields measurement outcomes. This overlap serves to bridge any potential semantic gaps between the manually modelled and the automatically generated ontologies, thereby facilitating a more accurate and meaningful comparison.

Overall, these design choices reflect a deliberate strategy to maximize the golden ontology's utility, while enabling a fair and productive comparison with the automated ontologies. These features are intended to ensure that the golden ontology serves not only as a robust knowledge representation tool but also as a useful benchmark for assessing automated ontology construction techniques.

4.2 Ontology Automation Methods

LDA for Automated Ontology Creation When selecting an algorithm for the automated creation of ontologies, our primary objective was to find an efficient and effective method for extracting topic information from unlabelled data. Given the extensive volume of the dataset and the constraints on resources, manual labeling of data was deemed impractical. Therefore, an unsupervised learning method was essential for this task, and Latent Dirichlet Allocation (LDA) was identified as the most suitable choice.
LDA is a widely used and well-established generative probabilistic model in the field of topic modeling as stated by Moghaddam et al. [13]. It is particularly adept at extracting latent topics (in this case, measurement outcomes) from large volumes of unstructured text data in an unsupervised environment. LDA achieves this by assigning topics to documents and words to topics, with the understanding that each document is a mixture of various topics, and each topic is a mixture of various words.

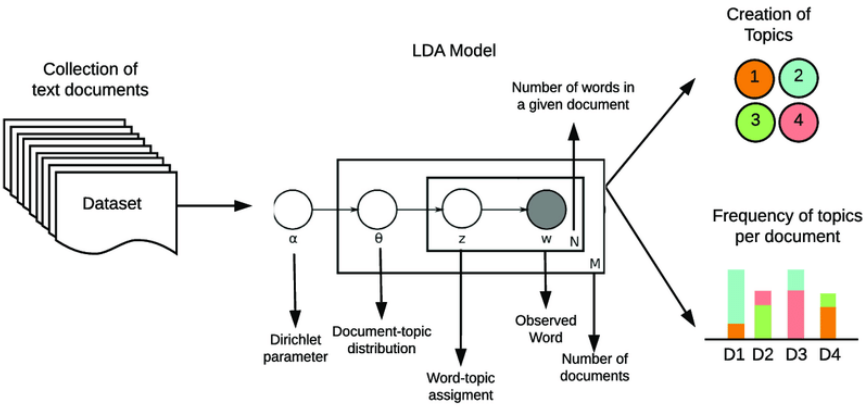


Fig. 2. Schematic overview of LDA algorithm [3]

Therefore, the decision to utilize LDA in our automated ontology generation process was driven by the need for a method that can efficiently handle unlabelled data and has proven effectiveness in extracting meaningful topics from large text corpora. With LDA, we aim to get topic/outcome measurements from the set of clinical trials that we feed it.

TF-IDF vs BOW: To ensure a robust and comprehensive automatic ontology, two distinct implementations of the LDA algorithm were adopted - *Bag of Words* (BoW) and *Term Frequency-Inverse Document Frequency* (TF-IDF). Each approach offers unique strengths, contributing to a more nuanced understanding of

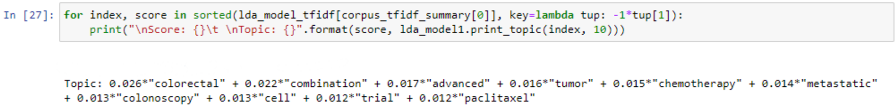
[3] https://www.researchgate.net/figure/Schematic-of-LDA-algorithm$_{fig}$1$_3$39368709

the dataset.

The BOW method is a straightforward approach, viewing each document as an unordered set of words, with a primary focus on their frequency. Despite its simplicity, this approach lacks the ability to capture the contextual relationship between words and can potentially overlook the significance of less frequent, yet potentially meaningful terms.
In contrast,the TF-IDF approach refines the process of assigning significance to terms within a document. It provides weights to words based on their prevalence within a particular document, as well as their distribution across the entire document corpus. By doing so, TF-IDF amplifies the importance of terms that appear frequently within specific documents but are sparse on a corpus-wide scale. Also rarer words are assigned greater weight, their scarcity across the document set suggesting a higher degree of relevance to the specific topics they occur in. For instance, a term like *elephant* in a clinical data context, would be infrequent, hence its presence may signify unique document-specific insights, even if semantically it would not be significant.
This strategic adoption and comparison of both BoW and TF-IDF approaches and results aim to generate a more sophisticated understanding of our data set. It enables a richer, more precise automatic ontology that can capture the intricacies of topics within the domain of clinical trials. Since within the GI cancers clinical data domain some words like ’cancer’ are much more common than ’safety’.

Measurement extraction through LDA:

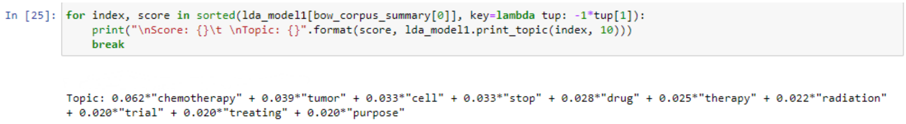
LDA models each document as a mix of a number of topics, where a topic is defined as a distribution over words. For instance, a paper may be 30 percent about *colorectal surgery* and 70 percent about *anastomosis procedures*. Each word in a document is probabilistically assigned to one of the topics. LDA then attempts to reverse-engineer this process, learning the potential topics and how much each document pertains to each topic.
The first document below depicts the code and the output for the topic/measurement extraction of the TF-IDF model for the first paper in our dataset.
The second picture depicts the same for our BOW model.



```
In [27]: for index, score in sorted(lda_model_tfidf[corpus_tfidf_summary[0]], key=lambda tup: -1*tup[1]):
             print("\nScore: {}\t \nTopic: {}".format(score, lda_model1.print_topic(index, 10)))
```

```
Topic: 0.026*"colorectal" + 0.022*"combination" + 0.017*"advanced" + 0.016*"tumor" + 0.015*"chemotherapy" + 0.014*"metastatic"
+ 0.013*"colonoscopy" + 0.013*"cell" + 0.012*"trial" + 0.012*"paclitaxel"
```

Fig. 3. Topic distribution of the first clinical trial in our dataset for the TF-IDF model



```
In [25]: for index, score in sorted(lda_model1[bow_corpus_summary[0]], key=lambda tup: -1*tup[1]):
             print("\nScore: {}\t \nTopic: {}".format(score, lda_model1.print_topic(index, 10)))
             break
```

```
Topic: 0.062*"chemotherapy" + 0.039*"tumor" + 0.033*"cell" + 0.033*"stop" + 0.028*"drug" + 0.025*"therapy" + 0.022*"radiation"
+ 0.020*"trial" + 0.020*"treating" + 0.020*"purpose"
```

Fig. 4. Topic distribution of the first clinical trial in our dataset for the BOW model

As seen, the distribution and the probabilities are different for both models, as their importance in their respective datasets are different. The outputs differ even if the input text is the same. This process shows that there are major differences in the coefficients that are assigned to each word for both the *TF-IDF*, and *BOW* models.

Reconstruction through OPEN AI's API:

After obtaining the specific topic distribution of each separate paper from our clinical trials, we have trained the *gpt-3.5-turbo-16k* model in order to get the desired output from a set of topics snippets. This translation, enabled by OpenAI's LLMs, is of significant relevance and is vital for the comparative analysis. Each class in the reconstructed format is systematically initiated with the term 'measurement', and subsequent found topics are appended accordingly. In addition, one of the significant advantages of implementing an LLM is the model's ability to interconnect various topics that may not possess individual significance but form coherent themes when collectively viewed.

For instance, consider the entity *measurement of gastric tumor growth*. The standalone term *measurement of growth* may not convey any substantial meaning. However, when used in consequent conjunction with the identified outcomes, it evolves into a complete entity. This semantic unity carries particular importance in the context of clinical trials for gastrointestinal cancers, where measurements like *measurement of gastric tumor growth* is a commonly used indicator.

The scheme below depicts the overall steps and processes that were undertaken for the reconstruction of the measurement extractions.

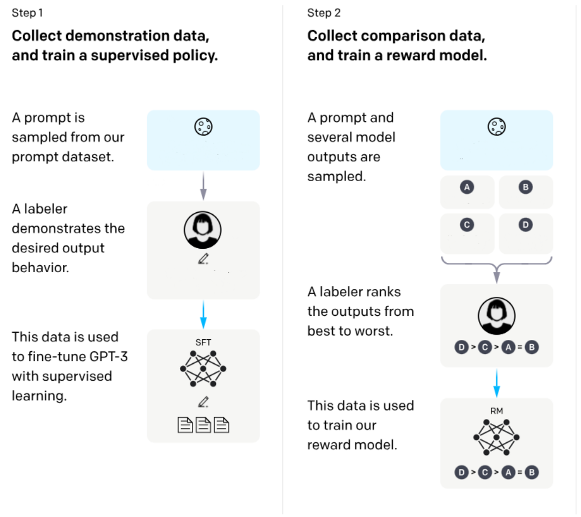

Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
A labeler demonstrates the desired output behavior.
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
A
B
C
D
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B

Fig. 5. Open AI's LLM training process [4]

From list to ontology: Owlready 2

In this research, the Python package Owlready2[5] was utilized. It provides a robust toolset for curating and generating ontologies. This package serves as the integral platform for processing and integrating the list of measurement outcomes produced by the OpenAI model. The workflow initiates with parsing the output list and subsequently removing duplicate entries to ensure the uniqueness of each measurement outcome. Following this data cleaning process, the Owlready2 package is called . The package allows for the creation of a new ontology and facilitates the addition of these cleaned and unique measurement outcomes. Owlready 2, with its capability to interface with the OWL (Web Ontology Language), serves as a pivotal component in the development and generation of the ontology.

5 Evaluation and Results:

The ensuing section covers the evaluative phase of the research. This stage underpins the empirical validity of the study, appraising the ontology generation processes that have been undertaken thus far. The structure of this evaluative segment is underpinned by a general overview of the resulting ontologies, and a comparative approach encompassing the results of the *TF-IDF* and *BOW* models scrutinized against each other.

[4] https://openai.com/research/instruction-following
[5] https://owlready2.readthedocs.io/en/latest/

5.1 Evaluation:

Initially, we will present the distribution of words within our clinical data corpus. This exploration of the total word distribution is a fundamental metric that provides us with critical insights into our dataset. By understanding the frequency of words, we can discern the prominence of various themes and topics within the clinical data corpus.

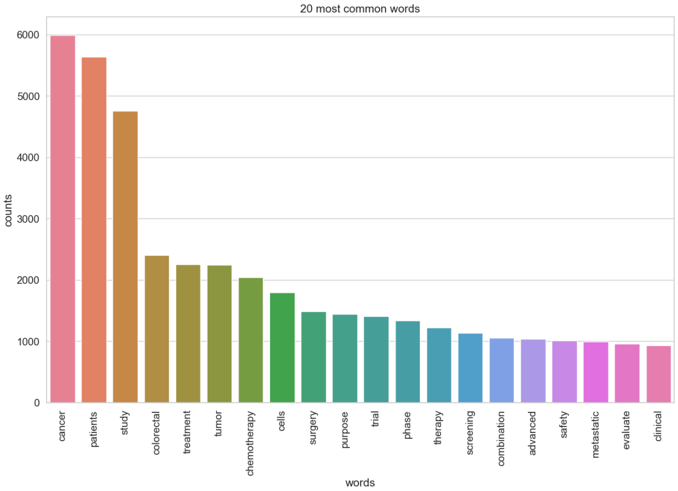


Fig. 6. Total frequency of words in corpus

Additionally, figure 7 in the appendix represent the number of classes identified within each ontology. These ontologies comprise the golden standard ontology, and those developed utilizing the *TF-IDF* and *BOW* models.

As seen the resulting *BOW* method has a total of 289 distinct classes, the *TF-IDF* model with a total of 477 classes, and the golden GI ontology has 659 classes. The difference between the number of classes between the two generated (*BOW* and *TF-IDF*) models can simply be explained by the difference in the coefficients assigned to each word in a document. The *TF-IDF* model assigns a seperate coeficient for each word in each document depending on the rarity of the word in question, wheras the *BOW* model maintains the same coeficient throughout the whole corpus of documents. This key disparity in assigning different importance to words results in the *TF-IDF* model having a larger pool of total measurements outcomes since more words are considered as crucial and not set aside because they only appear a few times in the whole set of corpus.
Furthermore, we delve into the comparison of the *BOW* and *TF-IDF* methods as shown on figure 9 in the appendix, specifically examining their 'absolute coverage', which denotes the commonality of items identified by both techniques.

The absolute coverage elucidates the degree of intersection between the ontologies generated by these two approaches, thereby revealing their shared capacity in recognizing and categorizing entities.

There is a total of 99 shared classes which indicate a level of consensus between the two methods, suggesting that these classes are likely to be particularly robust and reliably identifiable entities. Conversely, the unique classes — 190 for *BOW* and 378 for *TF-IDF* — reflect the distinctive characteristics of these two methodologies.The larger number of unique classes generated by the *TF-IDF* method hints at its potential for identifying a broader range of entities, possibly due to its emphasis on identifying terms that are important in a specific document, while down-weighting terms that are common across the entire corpus.

5.2 Results:

In the forthcoming section, we will present the results of this thesis, focusing particularly on high-level comparisons of all ontologies as seen on figure 9 in the appendix. This approach is favored due to the substantial size and complexity of the ontologies under consideration. With the Bag-of-Words (BoW) methodology yielding 289 classes, the Term Frequency-Inverse Document Frequency (TF-IDF) generating 477 classes, and the Golden standard ontology comprising more than 600 classes, it is evident that the ontologies are considerably extensive.

TF-IDF: In this section of the results, our primary focus lies on the performance of the TF-IDF model. Out of 16 main classes, the model has accurately mapped 10, partially mapped 2, leaving 4 main classes absent from the resultant TF-IDF ontology. These unmapped classes comprise *body morphological measurement*, *body temperature*, *consumption measurement*, and *organ measurement.*

The absence of organ measurement in the ontology could be attributed to the specificity of language typically employed in clinical trials. It is common practice to refer directly to the name of the organ being measured, rather than using the generic term *organ.* Consequently, the TF-IDF model, while picking up on the specific organ names, may overlook the overarching class of *organ measurement.*

Regarding 'body morphological measurement', this class might have been bypassed due to the granular nature of the discourse in most papers. For instance, more specific morphological measures such as 'muscle regression' or 'muscle hypertrophy' are typically mentioned, which might have led the TF-IDF model to assign these entities to more detailed classes, while missing the broader 'body morphological measurement' class.

The class 'consumption measurement' could have been overlooked for similar reasons. Clinical trials often describe consumption in the context of specific entities, such as 'sugar consumption' or 'glucose intake'. This level of detail could lead the TF-IDF model to map these concepts to more specific classes, thereby bypassing the overarching 'consumption measurement' class.

Intriguingly, the 'body temperature' class was not represented in the TF-IDF ontology, despite being frequently used within the provided dataset. This unanticipated omission eludes straightforward explanation and warrants further investigation into the specific behavior and limitations of the TF-IDF model in ontology generation.

BOW: In the subsequent portion of the results section, we direct our attention towards the ontology generated using the Bag-of-Words (BoW) model. From the 16 main classes, the BoW model correctly mapped 6 and partially mapped 2, leaving a substantial 8 main classes absent from the resultant BoW ontology. These overlooked classes encompass *Body morphological measurement*,*Body movement/balance measurement*,*Body temperature*, *Cardiovascular measurement*, *Consumption measurement*, *Musculoskeletal system measurement*, *Organ measurement*, and *Skin measurement* .
Analogous to the TF-IDF model, the BoW model missed *Body morphological measurement*, *Body temperature*,*Organ measurement*, and *Consumption measurement* classes, likely due to the reasons delineated in the prior discussion. Briefly, the specificity of the language in clinical trials and the granularity of the measurements described could have resulted in these classes being bypassed.
Turning our attention to the other classes which the BOW model did not map but that were part of the TF-IDF model, namely: *Body movement/balance measurement*,*Skin measurement*,*Musculoskeletal system measurement*,*Cardiovascular measurement*, it is plausible that these classes were not represented due to their inherent nuanced nature. In many clinical studies, the focus might lie on specific aspects of these classes, such as 'gait analysis', 'motor function', or 'Visceral control', thereby omitting the general term from the ontology.
Although if this were the case then the TF-IDF model would also fail to correctly map these classes. Another explanation may be that the BOW model did not find these measurement outcomes or classes important enough in the whole corpus of clinical trials and determined not to retain them as separate extracted topics.

5.3 Discussion:

In evaluating the performance of TF-IDF model against the BoW model, it is apparent that the former demonstrated a superior capability in terms of correctly mapping the main classes. Specifically, the TF-IDF model accurately identified 10 classes in comparison to the 6 accurately identified by the BoW model. Interestingly, both models had an equal performance in terms of partially mapped classes, with each model correctly identifying 2.
These results are consistent with the differences between the two models. The TF-IDF model, by design, accounts for the frequency of terms within each document relative to their occurrence across the entire corpus. In other words, it prioritizes terms that are common in a given document but less so across the corpus, thereby enabling it to discern nuances and specificities in individual documents. Consequently, this approach allows the TF-IDF model to accurately identify a variety of classes, even those that may be less common or more context-specific.

On the other hand, the BoW model operates on a simpler basis, assigning weights based on the total occurrence in the whole corpus. This means that the model treats each occurrence of a term in the corpus with the same importance. While this approach has its strengths, particularly in terms of simplicity and computational efficiency, it does not account for the contextual relevance and importance of terms. As a result, the BoW model may be less adept at identifying less common or more specific classes within a diverse and complex corpus.

6 Limitations and future work:

In assessing the merits and limitations of this research, it is crucial to acknowledge certain constraints for future enhancement. A limitation lies in the methodology of combining an unsupervised algorithm (LDA) with a Large Language Model (LLM) for ontology generation. While this approach has demonstrated efficacy in grouping and reshaping relevant words and measurements into coherent categories, it inherently does not consider hierarchical and associative relationships. Consequently, the ontology generated is primarily based on clusters of words or measurements, without the application of any further translations to refine the results. While this method yielded pertinent insights, it is not without potential areas for improvement.

Additionally, the evaluation process posed certain challenges. As it required manual examination of the major classes, and as the classes from the golden ontology did not perfectly coincide with those generated by the TF-IDF and BoW models, there exists a potential for human bias. Hence, this manual aspect of the evaluation process may introduce elements of subjectivity.

Furthermore this evaluation method does not take into account the quality of the automated ontologies, meaning that a portion of the created classes for both models might have been irrelevant. This is the case with the resulting ontologies because of the nature of the transformations applied to the corpus of clinical trials. Classes such as : *Measurement of like*, *Measurement of stop*, *Measurement of killing*, are relatively common in both models. Since this method does not take into account a checker for classes that are not pertinent, the model outputs any measurement reconstructed through OPEN AI's LLM .

Furthermore, there is considerable potential for future work on the LLM side of the methodology. The LLM used in this study was not extensively trained due to time constraints, which may have impacted its performance, specifically on the ontology quality aspect. Future research could focus on more extensive training of the LLM, potentially improving its ability to generate more accurate and coherent ontology classes, ultimately enhancing the overall effectiveness of the automatic ontology generation process. This additional training could also address some of the limitations encountered in the current study, including the handling of hierarchical and associative relationships.

Moving forward, future research could explore other methods of evaluation to more accurately assess the efficacy of this approach in automatic ontology generation. More sophisticated, automated or semi-automated evaluation strategies could provide a more objective and comprehensive understanding of the models' performance.

7 Conclusion:

In conclusion, this thesis explored the utility of Large Language Models (LLMs) such as OpenAI's GPT in combination with a Topic Extraction Algorithm (LDA) in programmatic ontology generation. This journey was driven by a central research question: How do ontology engineering approaches, which incorporate the usage of Large Language Models, compare in terms of quality to manually crafted ontologies?

Our findings indicate that the integration of LLMs into ontology engineering presents a promising avenue for progress, yielding respectable results, albeit with room for enhancement in terms of the quality of the resulting ontologies. Despite these observed limitations, the unique combination of Topic Extraction Algorithms (LDA) with LLMs, as demonstrated in this study, introduced a novel approach to programmatic ontology generation. While this method did manifest challenges in maintaining hierarchical relationships and ensuring the quality of generated ontologies, it represents a significant stride in the field of ontology generation, departing from traditional manual engineering methodologies which are often time-consuming and labor-intensive.

In light of these insights, future recommendations pivot around continued research to refine LLMs specifically for ontology generation methods. A focus on improving the model's handling of hierarchical relationships and enhancing the quality of the generated ontology could be fruitful areas of exploration. Further research could also investigate automated or semi-automated evaluation strategies to reduce potential bias in the assessment of ontology classes.

Reflecting on the contributions of this work, the thesis introduced an innovative, automated approach to ontology generation by combining LDA and LLMs, showcasing its potential in mitigating the challenges of manual ontology engineering. By reshaping the understanding and practices of ontology engineering, this research paves the way for future explorations into improved, efficient, and automatic methods of ontology generation.

8 Appendix:

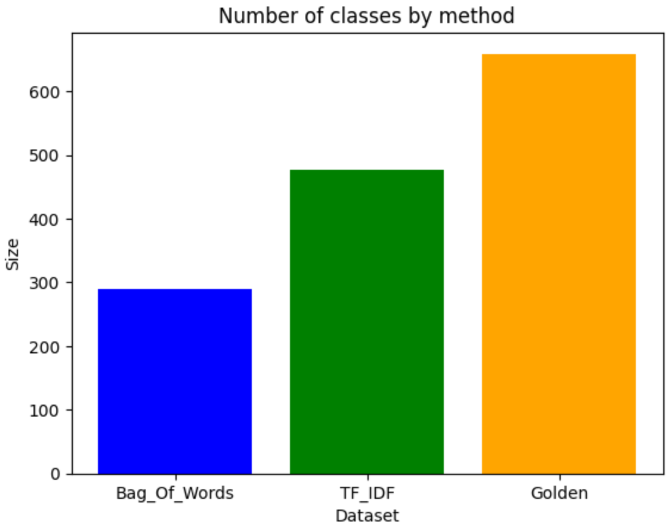


Fig. 7. The number of classes per ontology

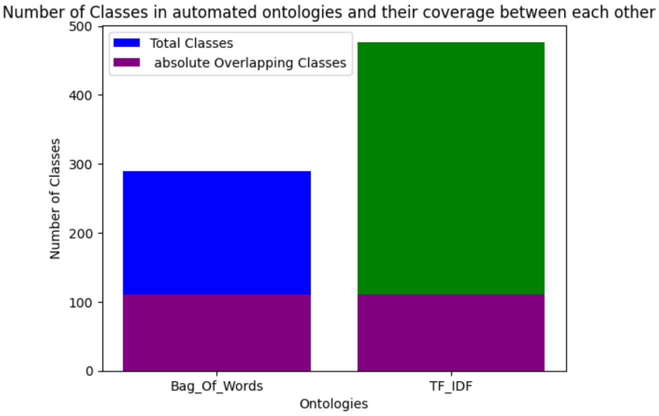
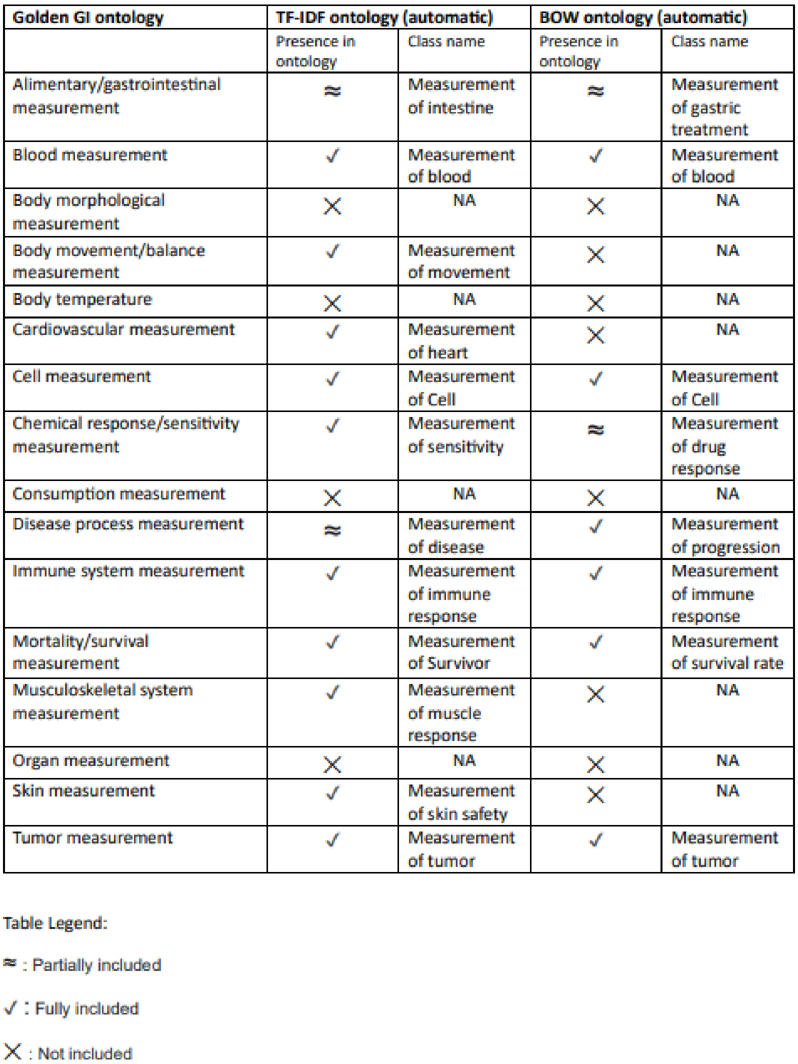


Fig. 8. Similarity between the *TF-IDF* and *BOW* models



Golden GI ontology	TF-IDF ontology (automatic)		BOW ontology (automatic)	
	Presence in ontology	Class name	Presence in ontology	Class name
Alimentary/gastrointestinal measurement	≈	Measurement of intestine	≈	Measurement of gastric treatment
Blood measurement	✓	Measurement of blood	✓	Measurement of blood
Body morphological measurement	✕	NA	✕	NA
Body movement/balance measurement	✓	Measurement of movement	✕	NA
Body temperature	✕	NA	✕	NA
Cardiovascular measurement	✓	Measurement of heart	✕	NA
Cell measurement	✓	Measurement of Cell	✓	Measurement of Cell
Chemical response/sensitivity measurement	✓	Measurement of sensitivity	≈	Measurement of drug response
Consumption measurement	✕	NA	✕	NA
Disease process measurement	≈	Measurement of disease	✓	Measurement of progression
Immune system measurement	✓	Measurement of immune response	✓	Measurement of immune response
Mortality/survival measurement	✓	Measurement of Survivor	✓	Measurement of survival rate
Musculoskeletal system measurement	✓	Measurement of muscle response	✕	NA
Organ measurement	✕	NA	✕	NA
Skin measurement	✓	Measurement of skin safety	✕	NA
Tumor measurement	✓	Measurement of tumor	✓	Measurement of tumor

Table Legend:

≈ : Partially included

✓ : Fully included

✕ : Not included

Fig. 9. high level coverage of classes with respect to golden GI ontology

Bibliography

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
2. OpenAI. Openai api documentation, 2021. URL `https://www.openai.com/api/`.
3. Chintan Patel, James Cimino, Julie Dolby, Achille Fokoue, Aditya Kalyanpur, Aaron Kershenbaum, ..., and Kavitha Srinivas. Matching patient records to clinical trials using ontologies. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 816–829. Springer Berlin Heidelberg, 2007.
4. Ravi D. Shankar, Susana B. Martins, Martin O'Connor, David B. Parrish, and Amar K. Das. An ontology-based architecture for integration of clinical trials management applications. In *AMIA Annual Symposium Proceedings*, volume 2007, page 661. American Medical Informatics Association, 2007.
5. Holger Stenzhorn, Gerhard Weiler, Mathias Brochhausen, Fabian Schera, Vangelis Kritsotakis, Manolis Tsiknakis, ..., and Norbert Graf. The obtima system – ontology-based managing of clinical trials. In *MEDINFO 2010*, pages 1090–1094. IOS Press, 2010.
6. Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. Exact: Automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making*, 10:1–17, 2010.
7. R. Sahay, D. Ntalaperas, E. Kamateri, P. Hasapis, O. D. Beyan, M. P. F. Strippoli, and S. Decker. An ontology for clinical trial data integration. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3244–3250. IEEE, 2013.
8. Wei Lin, Paul Babyn, and Wen Zhang. Context-based ontology modelling for database: Enabling chatgpt for semantic database management. *arXiv preprint arXiv:2303.07351*, 2023.
9. Hua Liu, Simona Carini, Zhaohui Chen, Spencer P. Hey, Ida Sim, and Chunhua Weng. Ontology-based categorization of clinical studies by their conditions. *Journal of Biomedical Informatics*, 135:104235, 2022.
10. David Lonsdale, Ying Ding, David W. Embley, and Alan Melby. Peppering knowledge sources with salt: Boosting conceptual content for ontology generation. In *Proceedings of the AAAI Workshop on Semantic Web Meets Language Resources*, July 2002.
11. Latifur Khan and Feng Luo. Ontology construction for information selection. In *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02)*, page 122, Washington, DC, USA, 2002. IEEE Computer Society.
12. Ilaria Bedini and Bach Nguyen. Automatic ontology generation: State of the art. Technical report, PRiSM Laboratory Technical Report, University of Versailles, 2007.

13. S. Moghaddam and M. Ester. On the design of lda models for aspect-based opinion mining. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 803–812. ACM, 2012.