Building and Enriching an Ontology on the basis of a Labeled Corpora of Opinions

Wojciech Waloszek
Gdansk University of Technology
Gdansk, Poland *wojwalos@pg.edu.pl*

Agnieszka Pluwak
SentiOne SA
Gdansk, Poland *agnieszka.pluwak@senti1.com*

Abstract

This paper presents a methodology for building and enriching an ontology from opinionated text, developed in collaboration between industry and academia. The focus of this work is on semantic alignment with Wikidata. Key contributions include a leading category-based scoring and approach and LLM-assisted refinement. Experimental results show that our leading category-based approach significantly improved alignment accuracy, reaching 86.5%. Furthermore, the incorporation of LLM-based refinement further increased accuracy to 90.6%, indicating the potential of this approach for automated ontology enrichment.

Keywords: ai-supported artifact generation, ontologies, concept hierarchies.

1. Introduction

In this paper we present results of a several month project conducted in collaboration between industry and academia. Our industrial partner, SentiOne, is a large company operating in several European countries and offering services for broadly understood sentiment analysis.

The central aim of this research project was to evaluate the utility of emerging artificial intelligence (AI) methodologies for sentiment analysis and opinion mining. Our investigation focused on determining the extent to which these advanced techniques can be leveraged to effectively monitor and interpret user feedback concerning products, ultimately providing valuable insights for businesses and product development. This includes understanding both the positive and negative sentiments expressed by users, as well as identifying key trends and areas for improvement.

One of the key challenges within the overall project lay in the integration of data from various sources and in multiple languages in order to build a knowledge base about products, their aspects and features [10] in a form of an ontology. This ontology should then be augmented with one of the knowledge graphs available in the Internet [3]. Our team focused on this aspect, building upon a pre-existing corpus of textual user reviews expressed in three different languages and spanned a range of domains, including, but not limited to, banking, health insurance, and technology (the dataset had been acquired from the internet). Crucially, this corpus had been tagged by a separate team of linguists, who dedicated several months to labeling the data according to a pre-defined schema. While their tagging provided a valuable foundation, the primary focus of our work was to develop means for integrating the data, augmenting them with missing semantic relations, and providing the end-user with tools for querying the data in various ways. This included enabling querying according to predetermined use scenarios, and facilitating the gathering of various aspects of sentiment across all languages.

For data integration, we employed an OWL ontology [1], constructed following a well-established ontology development lifecycle. The quality of the developed ontology was evaluated using the FOCA [2] methodology. To enrich the hierarchical relationships between concepts within the integrated data, we leveraged semantic data sourced from Wikidata [16]. The resulting unified data was represented as an RDF graph, enabling flexible querying using any compatible RDF engine. Furthermore, we conducted experiments to explore the potential of Large Language Models (LLMs) [17] for improving the process of concept hierarchy enrichment. This enrichment primarily addressed gaps in the initial corpus where, according to their set of predefined assumptions, the linguists had omitted numerous “obvious” semantic relationships (e.g., that an “antibiotic” is a type of “drug”), delegating this more

scalable task to our team.

The following sections of this paper present the background for, and the results of, our work. Section 2 provides a detailed characterization of the dataset. Section 3 outlines our ontology creation process. Section 4 is devoted to concept hierarchy enrichment, leveraging traditional ontological tools like Wikidata. Section 5 details the results of our experiments with the use of LLMs. Section 6 discusses our findings (also providing a short review of related works), and Section 7 summarizes the paper and presents our conclusions.

2. Dataset

The corpus consists of texts in German, English, and Spanish, sourced from a variety of online platforms including press articles, comments, forums, blogs, social media, and reviews. A significant portion of the corpus represents the colloquial register of computer-mediated communication. Semantically, the corpus covers the domains of healthcare, banking, and utilities (services), with a broadly similar distribution across languages.

The corpus was constructed using an internet monitoring tool with broad coverage across over 70 languages. This tool's crawlers allow for the creation of complex search queries using Boolean expressions. Source types included: portals, blogs, reviews, forums, and social media sites such as Facebook, Instagram, X (formerly Twitter), and YouTube. For this corpus, primary sources included Twitter, Facebook, Instagram, news services (e.g., www.dailymail.co.uk, https://finance.yahoo.com/), forums (e.g., forum.level1techs.com), blogs (e.g., blog.cloudera.com), and reviews (e.g., play.google.com). Data sources were proposed by SentiOne senior developers on the basis of their expertise, while queries were designed by domain experts, covering numerous sources and using brand and product names to expand the search, while also applying constraints to filter ambiguous keywords.

The resulting corpus comprises 18,077 English texts, 15,780 German texts, and 16,902 Spanish texts. Table 1 shows the distribution of internet source types across the three semantic domains. Reviews contribute the most texts to the corpus (13,752), followed by Twitter (9,378) and Forums (8,734). Facebook provides 8,041 texts, Portals 5,402, Blogs 3,141 and Instagram 2,311. The distribution of documents across the semantic domains is more equal, with 14,476 texts about Finance and Banking, 17,053 about Subscription Consumer Services, and 19,230 about Healthcare. Overall, the corpus contains 50,759 texts.

Table 1. Coverage of Internet Source Types in the Corpus.

	Finance and Banking	Services	Healthcare	Total
Forums	2901	2665	3168	8734
Facebook	2061	3178	2802	8041
Portals	1587	2087	1728	5402
Twitter	2209	3115	4054	9378
Reviews	4466	3442	5844	13752
Blogs	734	1443	964	3141
Instagram	518	1123	670	2311
Total	14476	17053	19230	50759

Annotation followed a 2+1 approach (two annotators and one senior annotator per language). Inter-annotator agreement (IAA) was evaluated on a sample set of 300 texts, representing all semantic domains and source types, using established methodologies.

During annotation, a rich set of labels was applied to the corpus. Primarily, these labels serve to identify:

- **Opinion Subjects**, encompassing a broad range of entities that may be evaluated by internet users. These subjects may include, but are not limited to: products, services, institutions, customer service, offers, and agreements.
- **Evaluation Aspects**, representing the characteristics through which opinion subjects are evaluated. These aspects may include, for example: speed, efficiency, location, usability, competence, reliability, price. In the case of aspects, a critical distinction is made between explicitly expressed aspects (e.g., “price”) and implicitly expressed aspects (e.g., “expensive”).
- **Positive or Negative Features** of aspects and opinion subjects expressed without reference to a specific aspect (e.g., good, super, cool, poor, mediocre, miserable).

In addition, several other types of labels were used in the corpus, particularly those for identifying specific products (**Product**) and product brands (**Brand**).

It should be noted that the actual structure of the labels in the corpus is somewhat more complex,

as it also accounts for situations where features are expressed in a distributed manner within a sentence (labels with suffixes 1 and 2) and information about the expression method and emotional valence of the text fragment. The labels and the original text have been combined into a single JSON file.

It is also worth noting, that while the initial corpus labeling was performed by a separate team of linguists, our respective work timelines overlapped, and our team was able to provide valuable feedback, contributing to the refinement of the subsequent corpus version. Specifically, using data visualizations in Neo4j and targeted queries in PostgreSQL, we identified inconsistencies and areas for improvement and communicated them back to the labeling team for consideration.

3. Ontology

Leveraging the tagged corpus as an initial structured resource, our team focused on development of an integrated data framework. Adhering to well-established principles of ontological engineering [1], we partitioned the ontological development process into two distinct sub-tasks:

- **Development of the Terminological Part (Class Hierarchy Specification)**: This involves the formal specification of a class hierarchy that models the key concepts and relationships relevant to the domain. The goal is to create a controlled vocabulary that serves as the foundation for semantic interpretation and reasoning.
- **Development of a Method for Ontology Population**: This focuses on the design and implementation of a semi-automated process for instantiating the ontology with data derived from the tagged corpus. The methodology aims to generate assertions that represent factual statements about specific entities and their interrelations, thereby creating a knowledge graph that reflects the information encoded in the original corpus.

In this section we will focus on the first sub-task, having in mind the previously stated primary objectives for the framework (namely: (1) facilitate the access to data, (2) augment the data with missing semantic relations, and (3) provide users with tools for more complex querying; all while supporting use-case specific information retrieval and the aggregation of sentiment aspects across multiple languages).

3.1. Development cycle

The development process for the terminological part of the ontology was conducted following a development methodology. Several such methodologies exist, varying in their level of maturity. For the purpose of this project the Methontology methodology [4] was adopted.

This choice was motivated by the following characteristics of Methontology:

- It is based on traditional software development life cycles, making it well-suited for creating ontologies of small to medium size.
- Its primary emphasis is on knowledge acquisition and conceptualization (whereas, for example, the NeOn methodology [13] primarily emphasizes reuse, and On-To-Knowledge [14] focuses on intensive communication with domain experts).
- It emphasizes the importance of ontology maintenance for future use, thereby increasing its utility for the end-user.
- The ontology development process can be tailored to specific needs.

Methontology is a highly structured ontology development methodology. It divides the ontology creation process into phases. These phases include specification, knowledge acquisition, conceptualization, integration, implementation, evaluation, and documentation.

3.2. Specification and Knowledge Acquisition

These phases of the Methontology process involve the identification of the ontology's purpose and scope, as well as the acquisition of the knowledge necessary for its construction. These two tasks were performed concurrently.

The knowledge required for constructing the ontology was primarily acquired through unstructured interviews with experts from our industrial partner, as well as through the study of documents pertaining to the presentation and aggregation of opinions for business clients.

Also usually during these phases, objectives are further detailed by introducing preliminary competency questions for the ontology. In this case, this was achieved by eliciting use-case scenarios; four main scenarios were identified:

- S1. Grouping opinion subjects by linguistic equivalents (e.g., monitor, Bildshirm, pantalla) along with synonyms (like: monitor, screen, display) to provide the client with an overview of opinions about a given subjects written in multiple languages/countries.
- S2. Connecting data levels by semantically grouping opinion subjects, enabling, for example, the evaluation of a bank based on its products. Conversely, the evaluation could also be

presented separately (flexibility in the levels of aggregation).

- S3. Opinion subjects are referred to by users in generic ways (telephone) and specific ways (Galaxy S5). The ontology should facilitate the grouping of generic and specific ways of referring to opinion subjects.
- S4. The ontology should enable the grouping of synonymous aspects, particularly those expressed explicitly and those expressed implicitly.

As can be noticed, these requirements focus on the semantic grouping of individual concepts identified during the corpus creation phase. The conceptualization presented in the following section of this article was proposed in this spirit.

3.3. *Conceptualization and Integration*

During the conceptualization phase, the main classes of the ontology were identified based on the knowledge acquired in the preceding stages. In this project, conceptualization followed a middle-out approach, meaning that the process began by identifying the most important classes and then progressively expanding the model.

Concurrent with the conceptualization phase, an integration phase was conducted to connect the developing ontology with existing ontologies. This connection was a key requirement, driven by the need to supplement ontological relationships present in the corpus. The original corpus labeling did not include relationships for objects higher in the hierarchy, i.e., more general knowledge (e.g., that a bank is an organization). Furthermore, it enabled the utilization of existing knowledge graphs to discover synonymous entities.

In the conceptualization of the ontology, the main classes were created in close relation to the three fundamental families of labels: Subject, Aspect, and Feature. The ontology simplifies the label structure, representing these three main classes linked by appropriate properties (see Fig. 1a).

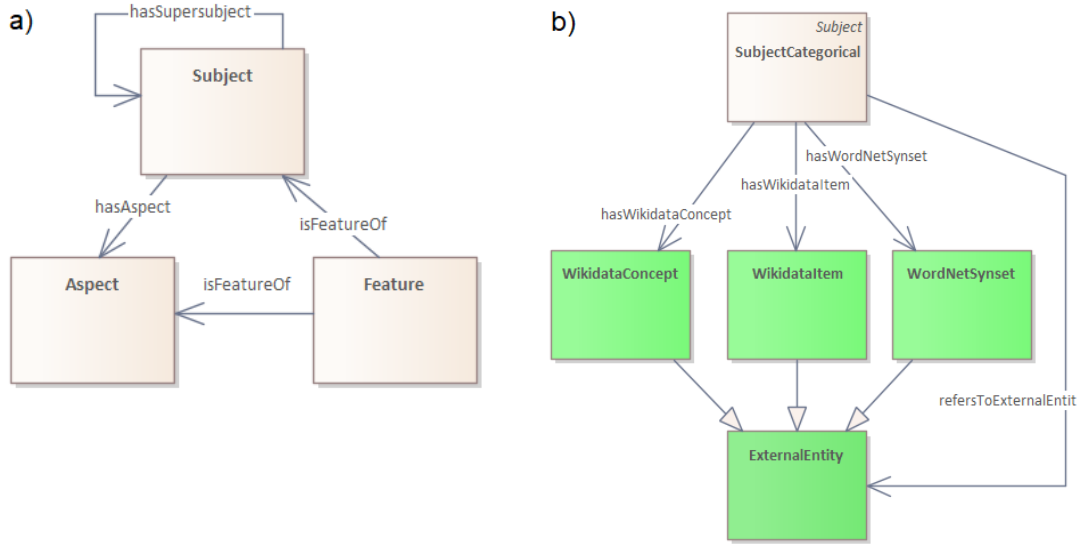

Fig. 1. Key concepts in the ontology (a) top level concepts, (b) concepts binding the corpus subjects to Wikidata items and concepts.

The *hasAspect* and *isFeatureOf* properties were derived from the corpus data. An additional property, *hasSupersubject*, was defined for the *Subject* class to reflect the hierarchical grouping of opinion subjects within the corpus and the ability to group such subjects, as detailed in scenario S2.

Given the central role of opinion subjects in sentiment analysis, the class hierarchy for these subjects was expanded along two dimensions. The first dimension classifies opinion subjects identified in the corpus as either “categorical” subjects, which can be placed within a specific hierarchical conceptual system (see scenario S2), or as "non-categorical" subjects. Non-categorical subjects are those whose meaning is difficult to assign to a specific domain without the context of another opinion subject (e.g., “100mg” as part of “Aspirin 100mg”).

The second dimension along which the class hierarchy for opinion subjects was expanded relates to the special significance of certain objects identified in the corpus. Currently, this primarily concerns products and their brands, but it represents a potential expansion point for the ontology in the future. These special types of opinion subjects are introduced as subclasses of the *Subject* class: *Brand* and *Product*.

A further challenge related to these phases was determining the form of integration with existing ontologies. Specifications discussed during the specification phase indicated that ontological links from existing knowledge graphs should enrich the concept hierarchy information contained within the corpus to provide improved responses in scenarios S1-S4. The integration issue was therefore twofold. In addition to the standard integration, which consists of relating the created ontology to existing ontologies in terms of classes and properties, the goal was to further enrich the information about opinion subjects.

Regarding the standard integration problem, a decision was made not to utilize upper-level ontologies [6]. This was due to the rather specific nature of the ontology, which, in a standard classification, could be considered an application ontology, as it aims to closely reflect the structure of a specific labeled corpus of documents. Potential integration with similar ontologies could disrupt this character and complicate its use. Nevertheless, a brief study was conducted to search for similar ontologies in the Linked Open Vocabularies (LOV) repository, without finding any in that database.

Semantic enrichment of the information contained in the corpus, on the other hand, was from the beginning one of the fundamental aspects of this task. Due to the potentially very wide range of domains to which the opinions relate, the use of very general knowledge graphs such as *DBPedia* (*Wikipedia*), *WordNet*, *YAGO*, *schema.org*, or *Wikidata* was considered from the beginning. Ultimately, the choice was dictated by technical considerations (see the next part of the paper). The connection was originally modeled in the ontology in a general way through a relationship (*refersToExternalEntity*) with an external entity (*ExternalEntity*), which was then refined to three derived relationships with the three finally used types of connections (with a *Wikidata* entry—*WikidataItem*, with the *schema.org* and *DBPedia* ontologies—via *Wikidata*, the *WikidataConcept* class, and with *WordNet*). Due to the non-categorical nature of the remaining types of opinion subjects, these external entities were connected with relationships to the *SubjectCategorical* class (see Fig. 1b). A similar schema of grouping has been employed for aspects, resulting in covering the needs for executing the scenarios identified during previous phases.

3.4. Implementation and Evaluation

The ontology implementation was performed using the OWL 2 language and the *Protégé* tool, version 5.5 [9], maintained by the University of Manchester and Stanford University. As part of this task, a study of ontology editors was conducted, considering a comprehensive set of tools including *Protégé*, *WebProtégé*, *Mobi*, *TopBraid Composer*, *Ontopic Studio*, *PoolParty*, *TopBraid EDG*, *Fluent Editor*, and the discontinued *Apollo*, *Altova Semantic Works*, *Hozo*, and *OwlGrEd*. This study led to the selection of Protégé due to its suitability for the project's constraints, namely a single principal ontology author and an ontology of small to medium complexity. The resulting ontology comprised 23 classes and 13 properties, exhibited $\mathcal{ALUH}(\mathcal{D})$ expressiveness [1], and was encoded in RDF/XML format.

The ontology was evaluated using the FOCA methodology [2], which is based on the Goal/Quality/Metrics (GQM) principle. The evaluation was conducted through a set of questions pertaining to the ontology, which were answered using a {0, 25, 50, 75, 100} scale, akin to a Likert scale.

Only the question concerning the utilization of other ontologies, the third question on the original list, received a score other than “100” (“50”). This reflects the lack of direct incorporation of higher-level ontologies. Nevertheless, the overall score, calculated using the formula provided within the FOCA methodology, is notably high, reaching 99.8%.

(It is also worth noting at this point that the lack of incorporating a higher-level ontology is mostly of the formal character here, and does not hinder the ability to combine two corpus ontologies with use of Wikidata concepts. Such combination can be still easily done with use of *WikidataItem* and *WikidataConcept* URIs, as they are defined by Wikidata service and have to stay the same between corpora. As such they can be rightly treated as higher-level concepts, which is the justification for “50” score. The lack of a higher-level ontology import means instead that the concepts like *Subject* or *Aspect* itself are not necessarily sufficiently explained in terms of such an ontology as—for instance—perdurants referring to use of such a subject or aspect in an internet review or opinion.)

4. Populating the Ontology

Populating the ontology was a crucial point of data integration within our work. This phase involved a synergy between importing data directly from our corpus, and then, augmenting and enriching it by incorporating information from an external knowledge graph. This approach allowed us to comprehensively represent the complete hierarchical structure of concepts across the domains of interest within our ontology, resulting in a richer knowledge representation, directly applicable to the identified scenarios.

The data import process was primarily a technical task. The source corpus, in JSON format, was imported into a PostgreSQL 11 database, which natively supports JSON data storage. Ontological objects were instantiated based on labeled text segments annotated within the corpus, and relationships between these objects were established based on dependencies identified within individual opinion texts.

In contrast, the task of enriching the ontology with information from a knowledge graph was significantly more ambitious, but also essential. This necessity arose from the inherent limitation that linguists could only identify relationships occurring within the scope of individual opinions. For example, the subject "Antibiotikum" (German term for antibiotic) never co-occurred with the term "Arzneimittel" (German term for medication). Without integrating and completing the conceptual hierarchy within the ontology, we would be unable to apply evaluation aspects related to medications to antibiotics for the German portion of the corpus.

4.1. Augmenting the Data with WikiData

For ontology enrichment, we elected to utilize the WikiData service and knowledge graph. The multilingual and multi-domain nature of our corpus immediately narrowed our options to a limited number of existing knowledge graphs with such characteristics, including DBPedia (derived from Wikipedia/the YAGO ontology), the multi-domain schema.org, and WordNet. Among these, we selected Wikidata, as the most modern, regularly maintained, and including references to both WordNet and DBPedia.

However, to perform the enrichment, the ontology and WikiData knowledge graphs first required alignment, primarily through semantic matching of concepts—specifically, subjects from the corpus and ontology to WikiData items.

For the preliminary alignment we utilized the entity search service provided by the Wikidata portal. This service requires the specification of a search term and language, and in response, it returns a ranked list of potential matches, where the ranking reflects the position in the search results (with 1 being the highest position, and subsequent matches numbered 2, 3, etc.).

For obvious reasons, the top match is not always the most appropriate. This is also due to the fact that only a single word, devoid of the context of the entire message, is passed to the Wikidata service. An illustrative example is the matching of the term “Santander”, which, in an unmodified search process, is associated with the city, whereas in our corpus, it refers to a bank.

To address the aforementioned issue, we introduced the concept of leading categories for a given corpus. The Wikidata knowledge graph contains items representing object classes. These classes are connected hierarchically using the special property P279 (subclass of) and to objects with P31 (instance of). For the purpose of describing our method, we define a category as a Wikidata item representing a class located sufficiently high in the hierarchy, such that its parent class is a general class (encompassing all objects). Leading categories are then defined as the categories most frequently assigned to subjects of type *Brand* and *Product*.

This approach is justified by the fact that the corpus is focused on identifying sentiment towards products and brands. Consequently, products and brands are the text segments most likely to reveal the specific terminology used within the corpus (e.g., the cosmetics industry, banking sector, medicine, consumer electronics products).

Therefore, to accomplish the task of selecting the appropriate Wikidata item for a specific subject from the corpus, two factors were ultimately considered: (1) the item's position in the ranked list returned by the entity search service, and (2) the leading category assigned to that item. These factors were incorporated through derivation of a heuristic formula for scoring individual item entries:

$$\frac{10-p}{10} + 0.08\sqrt{\frac{c}{\bar{c}}} \tag{1}$$

The first part of the formula simply considers the (zero-based) position p of the item in the ranked list returned by the entity search service. This form of the first addend converts this position to a score, highest of which can be 1 and lowest zero (we cut off the positions below 10th). Thus the position in the ranked list is treated as a linearly decreasing component of the sum (i.e., a lower position results in a lower score).

The second addend in turn, represents how “popular” among the brands and products of the corpus the leading category of the Wikidata item is. It is measured as the proportion of c, denoting the number of brands and products matched to the leading category, and $\bar{c}$, represents the average number of brands and products assigned to each leading category. However, our observation was that that the most popular leading categories had an order of magnitude higher number products and brands within them. Therefore, to reduce the score disparity between the most and less popular leading categories a square root operation was applied to the second component of the sum. The weight factor of 0.08 was then

chosen to ensure that the score boost coming from leading categories will not be larger than five ranking positions.

Since Wikidata items are language-agnostic, they might be treated as descriptions of subject semantics. The entity search service is in turn multilingual, so it can (and would) match the same language-agnostic Wikidata item to subjects being linguistic equivalents (like monitor, Bildshirm, pantalla), immediately fulfilling the requirements of scenario S1. Subsequently, the import of classes from WikiData, to form a coherent hierarchical tree, allowed for the completion of the ontology and addressing the remaining identified scenario requirements.

4.2. Implementation and Evaluation

The process of alignment and ontology completion was implemented using scripting languages (JavaScript). The resulting ontology, along with example RDF queries (illustrating solutions to the aforementioned issues involving antibiotics and Santander), were incorporated into a workbench, which was then shared with other teams for feedback and evaluation. The solution was very well-received.

An additional component of the workbench consisted of queries allowing for sample-based verification of the performed alignments. This sample-based approach was chosen due to the large size of the corpus, as the verification required work of people fluent in all three languages, effectively being a chokepoint for the task. (The queries, with use of RAND() SPARQL function [12], returned a sample of subjects with alignments, embracing about 5% of subjects for each language, which were then assessed by a reading person. It allowed us to substantially reduce the resources needed for the evaluation.)

Analysis of the initial version of the integrated ontology revealed an alignment accuracy of approximately 75.9%, but also identified a systematic error. Specifically, within the healthcare domain, drugs and chemical substances were frequently aligned not to WikiData items describing them, but to articles about them. The consequence of this was, of course, the incorrect assignment of categories (e.g., *Journal* instead of *Drug*).

We addressed these findings in a subsequent version of the ontology by introducing an additional mechanism into the scripts to detect misaligned publications and correct the leading category. This additional mechanism resulted in an improved alignment accuracy (again measured using a sample-based method) of 86.5%.

5. Using LLMs to Improve Results

To investigate the potential for leveraging emerging technologies to further improve the results, we conducted an additional experiment. This experiment utilized a large language model (LLM) to refine the meaning of individual subjects within the labeled texts.

Due to certain project constraints, we opted to use a local version of an LLM capable of running on a relatively standard desktop computer. The hardware we employed consisted of an AMD Ryzen 7 processor with 32GB of RAM and an RTX 4070S graphics card equipped with 12GB of RAM. This configuration enabled us to use a distilled version of the Phi-4 model (Microsoft) with 7B parameters. (The choice of a single local LLM was here justified by the specifics of the project, in which we aimed to answer a very general research question, whether use of a local LLM by a single engineer can improve the accuracy of ontology augmenting.)

After conducting tests on a small number of samples from the corpus and exploring various prompt formulations, we settled on a simple zero-shot query: “In what meaning has the word subject been used in this text? Answer shortly using only a few words.” This was followed by a simple query in the same context: “Give me a general concept for this meaning (max. 3 words).”

This process (which took approx. 19.5 hours in our hardware configuration) yielded a collection of concise statements regarding the meaning of individual subjects within their respective contexts, such as: “In this text, *Santander* refers to a financial institution, specifically a bank where the user previously held a current account. General Concept: Bank Name Usage.” While informative, these sentences could not be readily utilized to improve the alignment process. This limitation stemmed from the characteristics of the WikiData entity search engine, which performed optimally only when provided with a maximum of three words.

However, we identified specific cases in which we were able to leverage the results. These cases were:

- Subjects with a very general and/or diverse meaning, such as “digit.” In these instances, our scoring formula proved insufficient, as the proper contextual meaning received a very low rank in the search results. In these cases, a match between a category and the general concept identified by the LLM was used instead.
- Highly specific subjects with an empty list of matches returned by the search engine, an

example being “lady who took my order.” In these instances, the original subject was replaced with the general concept identified by the LLM (*Customer Service Representative*).

Together those changes allowed for achieving the accuracy of 90.6% (evaluated in the same way as described in Subsection 4.2).

6. Discussion

The topic of this paper lies at the intersection of several rapidly evolving research domains. The first is the use of Wikidata, extensively explored in [3], particularly for addressing multilingual challenges [15] or engineering problems [11]. The second is the application of large language models (LLMs) to ontology alignment [5], as demonstrated in works such as [8] and [7].

While both of these directions are actively being investigated, our approach uniquely combines these concepts into a relatively concise and lightweight framework characterized by:

- A strong emphasis on data semantics, particularly in our leading category-based scoring approach,
- An engineering-oriented design, integrating the alignment process as a fundamental aspect of terminology management and weaving it into the ontology development lifecycle.
- A resource-efficient (yet impactful) application of LLMs, focusing on a targeted (“smart”) rather than exhaustive (“heavy”) utilization of emerging technologies, with potential for further reduction by limiting LLM usage to particularly problematic cases. Notably, our approach achieves results comparable to those reported in [8], despite employing significantly less powerful hardware resources.

The approach has been successfully applied within our project for a corpus of documents embracing three main domains of interests and expressed in three languages. However, we believe that the generalized version of it might be beneficial for other projects. The diagram of this generalized version is presented in Fig. 2. A large subset of components created as a result of our project might be reused, and we included the names of the components in parentheses. The process is fairly linear, with the possible reiteration of LLM-supported augmentation which we expect more corpora-dependent.

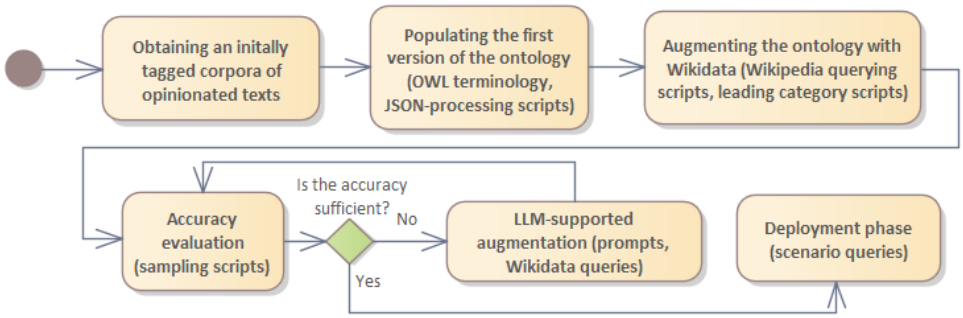

Fig. 2. A generalized version of the ontology augmentation and deployment process, in parentheses we included the components that are ready to be reused.

7. Conclusions

This paper addressed the challenge of building and automatically enriching an ontology from a multilingual text corpus of opinions, focusing on improving the accuracy of semantic alignment between corpus subjects and Wikidata items.

We presented a methodology for ontology enrichment that combines corpus-based analysis, Wikidata integration, and large language model assistance. Our key contributions include: (1) a leading category-based approach for improving Wikidata item alignment, (2) a demonstration of the effectiveness of light use of LLMs in resolving ambiguity in subject meaning, and (3) a refined ontology with improved accuracy and coverage of relevant concepts.

Experimental results show that our leading category-based approach significantly improved alignment accuracy, reaching 86.5%. Furthermore, the incorporation of LLM-based refinement further increased accuracy to 90.6%, indicating the potential of this approach for automated ontology enrichment.

The results might be impactful also because of the possibility of their reuse in other projects. Figure 2 presents a possible workflow, including the prepared components (in parentheses) which might be utilized to create a new augmented ontology from another corpus. This ontology can also be combined with other ontologies in a larger knowledge graph (see the remark in Subsection 3.4).

While our results are promising, the Reader has to remember that they have been obtained with use of a single corpus of data and still need to be generalized. For this purpose we plan to reuse the process

in our subsequent projects. In a similar fashion, the LLM part of the study was limited by the reliance on a specific LLM and a fixed set of prompt formulations. Future work should explore the use of different LLMs, investigate more sophisticated prompting strategies, and evaluate how our approach generalizes to other corpora and domains. Analyzing the types of errors that still occur could guide the refinement of the alignment process even further.

Acknowledgements

Research presented here has been partially funded by The National Centre for Research and Development in Poland, grant POIR.01.01.01-00-0923/20 (“SentiDeepFusion”).

References

1. Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., Patel-Schneider, P. F. (eds): The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, (2010)
2. Bandeira, J., Bittencourt, I., Espinheira, P., Isotani S.: FOCA: A Methodology for Ontology Evaluation. Applied Ontology, (3) (2015)
3. Farda-Sarbas M., Müller-Birn C.: Wikidata from a Research Perspective - A Systematic Mapping Study of Wikidata. CoRR (2019)
4. Fernández-López M., Gómez-Pérez A., Juristo N.: METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In: Proceedings of the Ontological Engineering AAAI-97 (1997)
5. Giglou, H. B., D'Souza, J., Karras, O., Auer, S.: OntoAligner: A Comprehensive Modular and Robust Python Toolkit for Ontology Alignment. ESWC 2025 (2025)
6. Guarino, N.: Formal Ontologies and Information Systems. In: Formal Ontology in Information Systems. Proceedings of FOIS’98 (1998)
7. He, Y., Chen, J., Dong, H., Horrocks, I.: Exploring large language models for ontology alignment. arXiv preprint arXiv:2309.07172 (2023)
8. Hertling, S., Paulheim, H.: Olala: Ontology matching with large language models. In: Proceedings of K-CAP ’23, ACM (2023)
9. Knublauch, H., Horridge, M., Musen, M., Rector, A., Stevens, R., Drummond, N., Lord, P., Noy, N., Seidenberg, J., Wang, H.: The Protege OWL Experience. In: Proc. of the Fourth International Semantic Web Conference (ISWC2005) (2005)
10. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool (2012)
11. Pfundner, A., Schönberg, T., Horn, J., Boyce, R.D., Samwald, M.: Utilizing the Wikidata system to improve the quality of medical content in Wikipedia in diverse languages. Journal of medical Internet research (2015)
12. Staab, S., Studer, R. (eds): Handbook on Ontologies. Springer, Berlin, Heidelberg, (2004)
13. Suárez-Figueroa, M. C., Gómez-Pérez, A., Fernández-López, M.: The NeOn Methodology for Ontology Engineering. Springer (2012)
14. Sure, Y., Staab, S., Studer, R.: On-To-Knowledge Methodology (OTKM). In: Handbook on Ontologies, pp. 117–132, Springer Berlin Heidelberg (2004)
15. Turki, H., Vrandečić, D., Hamdi, H., Adel, I.: Using WikiData as a Multilingual Multi-dialectal Dictionary for Arabic Dialects. In: 14th Intl. Conference on Computer Systems and Applications (2017)
16. Vrandečić, D.: Wikidata: a new platform for collaborative data collection. In: 21st International Conference on World Wide Web, pp. 1063–1064, ACM (2012)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems 30 (2017)