Journal of Computer Science

Original Research Paper

Enhancement of Semantic Analysis Based on the Ontology Reengineering: A Case Study on Moroccan Tourism Domain

[1]Khalid Tatane, [2]Asma Amalki and [2]Ali Bouzit

[1]National School of Applied Sciences, ESTIDMA Research Team, Ibn Zohr University, Agadir, Morocco
[2]Faculty of Science, Image and Pattern Recognition-Intelligent and Communicating Systems Laboratory (IRF-SIC), Ibn Zohr University, Agadir, Morocco

Article history
Received: 23-01-2024
Revised: 22-03-2024
Accepted: 18-04-2024

Corresponding Author:
Khalid Tatane
National School of Applied Sciences, ESTIDMA Research Team, Ibn Zohr University, Agadir, Morocco
Email: k.tatane@uiz.ac.ma

Abstract: This article outlines a novel semi-automatic approach aimed at enhancing a foundational ontology associated with Moroccan tourism. The initial phase of this methodology involved extracting new conceptual entities and semantic relationships from heterogeneous and multi-format textual resources, using a meticulously curated suite of Natural Language Processing (NLP) tools. The subsequent phase focused on aligning this updated Ontological version (OTMv1: Ontology of Tourism in Morocco version 1.0) with other external business ontologies. This was achieved through the design and development of specific terminological and structural matches to facilitate the mapping of knowledge shared by these ontological resources. The implementation of this approach resulted in the reengineering of an enhanced and semantically richer Ontology version (OTMv2). To streamline the validation tasks associated with the ontological model, particularly involving domain experts, we introduced a dedicated web platform (WebApp OTMv2) for experimentation and testing. Research conducted on this platform confirms the effectiveness of the proposed approach in enhancing the semantic quality of the outcomes within the context of Moroccan tourism.

Keywords: NLP Tools, Domain Sub Ontologies, Ontology Alignment, Ontology Enrichment, Ontology Reengineering

Introduction

Computer ontology plays a pivotal role in representing the knowledge hidden in texts and making it understandable by both humans and machines. Its construction provides various semantic solutions, including knowledge management, sharing, organization and enrichment (Pressat-Laffouilhère, 2023). Traditionally, these kinds of concept-semantic connections are built and maintained manually, or else the rate of human error would remain higher. Given the abundance of tourism-related information present on various resources, as well as the scale of the associated efforts and costs, the option under consideration is to fully or partially automate the laborious tasks involved in the process of creating and upgrading ontologies related to this domain. This article presents the phases and results of a so-called semi-automatic methodological approach aimed at extending an ontological kernel, OTMv1 (Ontology of Tourism in Morocco). The latter was initially developed manually by analyzing the semantics of a specialized thesaurus as part of a research project entitled knowledge management and WEb Semantic-GECO-WES (Bailal *et al*., 2023), initiated by the laboratory (IRF-SIC) of the faculty of science at Ibn Zohr University.

The research carried out in the context of developing the results of this preliminary study was deployed in two phases.

Phase 1: Analysis of texts and extraction of linguistic and semantic elements related to the domain in question, using language-processing tools (Tatane, 2023).

In general, it is accepted that the construction of a domain ontology based on the analysis of textual documents must go through the following stages: (i) Constitution of a corpus of documents, (ii) Linguistic analysis of the corpus, (iii) Semantic normalization and (iv) Formalization of the ontology.

Phase 2: Alignment of business sub-ontologies related to national tourism, using linguistic and structural matchers (Tatane, 2023).

The aim of this phase is, on the one hand, to combine the knowledge or, more precisely, the common concepts distributed in different sub-ontologies already identified (accommodation, transport, health, leisure, sport... etc.,). On the other hand, it aims to evolve the OTMv1 core, so that it is semantically richer, more expressive and better connected with other ontological resources (external and internal).

The involvement of domain experts in the first phase and business experts in the second is essential. The intervention of the cognitive scientists in the first phase consists of:

- Cleansing parasitic results
- Interpretation and validation of results obtained from corpus analysis

The second phase consists of:

- Validation of the new elements integrated into the predefined ontology
- Validation of correspondences between the ontology and other sub-ontologies
- Evaluation of the new, enriched ontology

Given that, domain and business specialists are not always available, an (online) web platform has been designed to facilitate access to them and enable them to carry out their tasks more simply and efficiently.

In this article, we have proposed a new semi-automatic approach to ontology enrichment based on terminological, relational and semantic analysis of a study corpus related to Moroccan tourism, as well as a process of mapping between the concepts of the base ontology and other internal and external business sub-ontologies identified.

However, the remainder of this article is structured as follows: Part 2 presents the general context of our study, followed by Part 3 describes related works, while Part 4 outlines the various treatments proposed at the approach level to evolve the base ontological core. Part 5 will be dedicated to validating the results and their experimentation. In the last two parts, we will provide a general conclusion along with some research directions that may guide our future interventions in this field.

General Context

Creation of a Representative Corpus of Domain

A corpus is a set of text fragments designed and compiled to reflect a particular variety of language or context of language use and to answer specific research questions (Clancy and Vaughan, 2023). Indeed, according to Sree Harissh *et al*. (2018), the domain corpus is a coherent collection of texts related to a domain.

Extraction of Candidate Terms

Terminology extraction is an essential task in the acquisition of domain knowledge and information retrieval. It is also an essential first step in building/enriching terminologies and ontologies (Lossio-Ventura *et al*., 2016). Terminology extraction, of which tracking and segmentation are the major operations, has been fully automated for decades and it has become absurd to process a specialized corpus without resorting to automatic extraction of its terminology or to a "term extractor" (Elbacha, 2023). Term extraction methods generally involve two main stages. The first stage consists of extracting units that are likely to be Candidate Terms (CTs), while the second stage consists of skimming and validating them by terminologists in order to decide on the terminological status of these candidate terms. Once validation is Complete, the (CT) list is transformed into the actual term list (Elbacha, 2023). However, different methods can be used for automatic term extraction (Elbacha, 2023):

- The linguistic method: Relies primarily on linguistic knowledge of the lexical, morphological, syntactic and/or morph syntactic order
- The statistical method: This relies exclusively on quantitative measurements and calculations of numerical values and, unlike the previous method, does without linguistic knowledge
- The hybrid or mixed method: This is generated as a natural result of the fusion of the two previous methods, to fill in the gaps and exploit the advantages

Lexical Relationship Extraction

Relationship extraction is an essential aspect of information extraction, aiming to discern the semantic relationship between pairs of entities presented in natural language text. These entities may be linked explicitly or implicitly (Gao and Liu, 2023). Relationship extraction is of paramount importance in diverse applications and its techniques have been widely applied in knowledge graphs, question-and-answer systems, information retrieval, intelligent customer service and various other fields (Zhu *et al*., 2023). Approaches for extracting lexical relations can be classified into two groups (Li and Mao, 2019):

- Approaches that use symbolic methods based on lexico-syntactic models, which are either elaborated manually or inferred automatically (Li and Mao, 2019)
- Purely statistical approaches present techniques for extracting term hierarchies based on word frequency and co-occurrence values (Wang *et al*., 2018)

Natural Language Processing

Natural language processing is a sub-field of computer science concerned with the intelligent processing of human language (Ford *et al*., 2016). It has had a major influence on documentary research and is still highly relevant in the field of terminology studies, particularly when it comes to using corpora (Nasr, 2023). With the rapid development of information and computer technologies, TALN has played an important role in various applications, including conversational agents and dialogue systems (Gašić *et al*., 2017), machine translation (Dai and Liu, 2024), knowledge extraction and reasoning (Moens, 2018), search engines (Zhou, 2018), etc. Although great progress has been made in the field of NLP, it remains a great challenge due to the inevitable ambiguity of representation in natural languages and the continuous evolution of vocabulary and syntax (Chen and Luo, 2019).

Domain Sub-Ontologies Alignment

Alignment is the process of identifying correspondences between entities belonging to different ontologies (Ngo and Bellahsene, 2016). It represents an essential solution to the challenge of semantic heterogeneity, by identifying possible correspondences between distant ontological entities (Ngo and Bellahsene, 2016). A correspondence is a triple of the form $\langle e1, 2, r \rangle$ with O_1 and O_2 two given ontologies, $e1 \in O_1$ *and* $e_2 \in O_2$ are elements of the ontologies to be mapped and r being the relationship between the two elements. Examples of relationships are equivalence ($\equiv$) or inclusion ($\sqsubseteq$). Correspondence may be accompanied by an explanation e and a confidence value c, it is sometimes described as a quintuple of the form $\langle e1, 2, r, c, e \rangle$. We distinguish two types of correspondences: Simple correspondences, which link an element of O_1 to an element of O_2 and complex mappings, i.e., mappings that contain logical constructors or transformation functions (Portisch *et al*., 2022).

A matcher is a mapping algorithm designed and developed to identify correspondences between ontology elements (Ngo and Bellahsene, 2016). An alignment system can be considered as a combination of three main components: (1) A terminology-based matcher, (2) A structure-based matcher and (3) A semantics-based matcher accompanied by a match selection module (Essayeh and Abed, 2015). Although these components exploit different characteristics of ontology entities, they are not independent of each other. A structure-based matcher takes as input the matches resulting from a terminology matcher (Essayeh and Abed, 2015) and a semantics-based matcher can take as input the matches resulting either from a terminology matcher, or a structural matcher, or a combination of the resulting matches of the two.

Related Work

In this section, we provide a brief overview of previous research work related to ontology extension from texts. This technique involves incorporating additional instances of concepts and relationships into an ontological resource. Typically, this intervention begins after the initial structure of the ontology has been developed. However, enriching this core ontology, whether automatically or semi-automatically, currently represents an area requiring further research. We emphasize that several approaches can be considered for ontology extension and as indicative researches, we may present:

- NLP-based approach: (Makki, 2017) has proposed a semi-automatic method based on NLP for extending risk management ontology from textual corpus. An automatic methodology was proposed by Labidi *et al*. (2017) based on NLP techniques for enriching the ontology of intentions from textual client requests in the IT market
- Ontology Mapping based approach: (Djellali, 2013) proposed a semi-automatic approach that uses the variables selection and clustering to find the candidate changes of an ontology, the approach uses an alignment process to find the rules of correspondence that define how to transform entities by defining all types of possible associations between ontological entities and candidate changes. (Cardellino *et al*., 2017) have suggested a prototype to align two ontologies of the legal domain, LKIF and YAGO. This approach has led to the enrichment of LKIF ontology

In this study, we proposed a semi-automatic approach that combines the two previous kinds of approaches, for extending a domain ontology and enhancing semantic analysis for applications based on ontological kernels. We have also applied the proposed methodology to the Moroccan tourism domain.

Proposed Methodological Approach for Upgrading the Basic Ontology Core (Otmv1)

Phase One: Semi-Automatic Ontology Kernel Enrichment Using Textual Resources

The semi-automatic enrichment approach for the Moroccan tourism ontology OTMv1, proposed in this first phase, aims at extending this basic core by integrating new concepts and semantic relations extracted from unstructured texts. Figure 1 illustrates the six stages that make up this approach as follows:

- Corpus creation: Creation of a domain-specific corpus from heterogeneous, multi-format text resources so as to take account of different conceptual-semantic viewpoints. The total number of words collected converges to 1 million words
- Corpus pre-processing: This phase consists of unifying text formats into a single raw format, using a unified UTF8 encoding, processing punctuation, abbreviations and acronyms, replacing unit symbols and several other operations linked to text standardization
- Corpus standardization: Using the two standardization techniques of automatic natural language processing, word lemmatization and sentence segmentation
- Linguistic analysis: A phase involving the morph syntactic labeling of texts, the extraction of candidate terms using linguistic, statistical and mixed approaches and the extraction of lexical relations using syntactic approaches or those based on external terminological resources
- Semantic standardization: This phase involves verifying the uniqueness and homogeneity of the concepts and semantic relations identified, with reference to the ontological components already presented in the original model
- OTMv1 enrichment: Through the integration of new concepts and semantic relations into the active OTMv1 core

Table 1, presented below, illustrates all the interventions carried out as part of this study, for the development of a new, semantically richer ontological product from various textual resources.

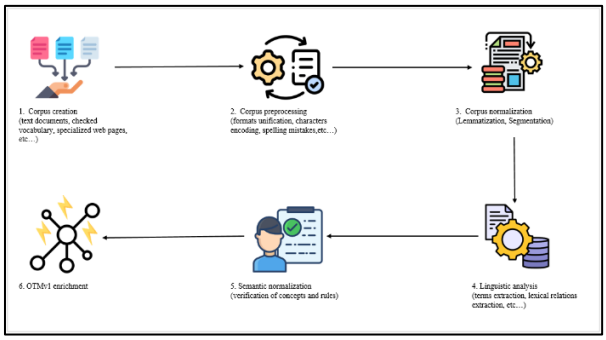

Fig. 1: Proposed approach for enriching the OTMv1 ontology kernel from textual resources

Table 1: Work carried out to upgrade OTMv1

Step	Description	Tools	Results	Degree of automation
Creation of Specialized corpus (data collection)	The corpus was created by and heterogeneous texts (blogs, web portals, text files, electronic newspapers and magazines, blogs, etc.) This was done to cover the maximum amount of knowledge distributed around Moroccan tourism The web resources used are official web portals of certain tourism organizations (public and private) Summary of knowledge sources used: 300 web pages, 10 PDF files, 8 PPT files, 10 tourism magazines, collected from the official websites of the following organizations: 1- Ministry of Tourism, Handicrafts and the social and Solidarity Economy of Morocco 2- High Commission for Planning of the Kingdom of Morocco Bellefleur, 2011), 3-UNWTO world tourism Barometer, 4- World Bank-tourism data	a Web crawler	Over a million words to analyze	Manuel
Corpus pre-processing	Unification of HTML, PDF and PPT formats into a single, native text format	Tools for converting to native formats	Reduce lexical inconsistencies and syntactic ambiguities	Semi-automatic

Table 1: Continue

	Application of unified UTF8 encoding Correction of spelling errors Standardization of case Punctuation handling, except for punctuation representing compound words or delimiting sentences Treatment of abbreviations and acronyms. Marking paragraph ends with spaces Treatment of numbers, conversion into textual form Treatment of unit symbols, conversion to text form			
Corpus standardization	Word lemmatization Sentence segmentation	Tree tagger	Reduce lexical inconsistencies and syntactic ambiguities	Automatic
Linguistic analysis	Morph syntactic tagging of texts using the tree tagger tool Extraction of candidate terms using the two-term extractors, YaTeA and TermoStat Analysis of YaTeA and TermoStat results, by calculating quality, precision, recall and Measure function indicators TermoStat is the term extractor, based on the previous analysis Extraction of conceptual relationships	-Tree tagger (French version) -YaTeA (Yet another term ExtrActor) -TermoStat -Gate-developer. -WOLF lexical database	Each word is unambiguously marked and assigned a lemma that defines its grammatical category	Automatic
Semantic standardization	Convert valid candidate terms into concepts Conversion of valid lexical relations into semantic conceptual relations Verification of concept definition by domain experts Verification of semantic relations definition by domain experts	_	Guarantee the uniqueness concepts and the homogeneity of concepts and semantic relations across the different hierarchical levels the ontology	Manuel
OTMv1 core enrichment	Integration of new concepts and semantic relations into of the OTMv1 ontology core using OWL Compliance and consistency analysis of the newly enriched ontological model	The FACT ++ reasoner of protege 2000	Ontology of tourism in Morocco enriched and updated with 20 new business concepts	Semi-automatic

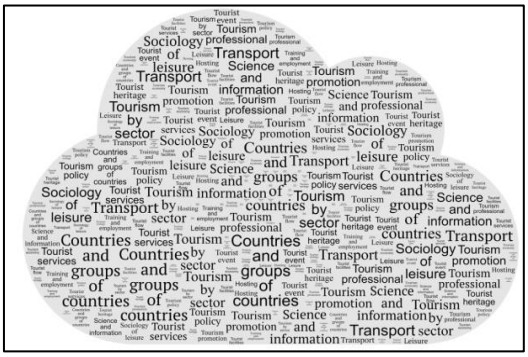

Fig. 2: New concepts integrated into the OTMv1 ontology core

The new OTMv2 ontology model, developed using the suggested methodological approach, features increased semantic richness. It is based on over twenty new generic business concepts, offering a broader semantic scope for future queries in the domain of national tourism.

The approach proposed in this phase enabled the semi-automatic enrichment of the basic OTMv1 ontology created as part of the GECO-WES project, by adding new concepts and relations as shown in Fig. 2. The process began with the creation of an expressive tourism corpus, encompassing a million entities to be analyzed. The application of NLP tools on this corpus enabled the extraction of 4412 candidate terms and 1114 lexical

relations. The standardization process then converted the candidate terms appropriate to the Moroccan tourism context into concepts and the lexical relations into semantic relations. Finally, after validating the results obtained, evaluating and integrating the new semantic entities into the basic ontological model, the whole process succeeded in extending the basic ontological kernel OTMv1, making it more expressive and semantically richer.

Analysis of this new version revealed that the new ontological model obtained is only a global ontology with direct connections to other ontologies (business) directly related to tourism-related sub-activities. Against this backdrop, new thinking has been initiated with a view to merging certain ontological models already developed, enriched and, above all, validated by other studies. More specifically, the aim is to match the ontological elements presented in two distinct yet convergent models. To meet this need, a new phase of study, focusing mainly on the alignment process, has been undertaken.

Second Phase: Aligning the OTMv1 Global Ontology with Support for Identified Business Sub-Ontologies

In the context of the GECO-WES project, the approach proposed in this phase complements that described above, still aimed at the semi-automatic extension of the tourism ontology in Morocco, through ontological alignment or mapping between conceptual entities belonging to the basic OTMv1 semantic network and those presented at the level of tourism-related sub-ontologies such as accommodation, transport, health, leisure, sport, etc. This study relies mainly on two types of ontological alignment techniques for calculating similarities:

- Terminology techniques involving string comparison
- Structural techniques including constraint processing and attribute analysis within concepts

The proposed methodological approach, illustrated in Fig. 3, begins with the selection of a source ontology, presented at the level of the structure of our field of study and a target ontology (the object of a mapping possibility). For each concept in the source ontology, if not already covered, a search for equivalent synonyms is carried out on a linguistic reference database (the Euro word-net lexical database). Calculations of terminological and structural similarities (equivalence, synonymy, subsumption, etc.,) between concepts in the source ontology and the target ontology are then carried out using the proposed alignment matchers. The results obtained are stored in a similarity database known as a temporary mapping database and then examined by business experts to guarantee semantic consistency. The validated similarity results are then injected into the final mapping database, to derive structural similarities using structural matchers and finally to guarantee consistent semantic matching results appropriate to the OTMv1 base ontology.

Terminological Matchers

Dictionary-Based Matcher

The aim of the dictionary-based matcher is to identify the different types of relationships between two concepts, shown below, by querying and manipulating the results obtained from the euro word-net lexical database, which provides synsets (a set of synonyms) for each concept. Types of relationships that can link two concepts C1 and C2 are:

- C1 ≡ C2 if there is a meaning for C1 synonym of C2
- C1 ⊇ C2 if there is a meaning for C1 hypernym of C2
- C1 ⊆ C2 if there is a meaning for C1 hyponym of C2
- C1 ⊥ C2 if there is no relation between meanings of C1 and C2

Matcher Based on Jaro-Winkler Distance

The Jaro-Winkler distance is a measure of similarity between two strings, used to establish links between entities, records and data cleansing (Wang *et al*., 2017). The higher the Jaro-Winkler distance value for two strings, the greater the similarity between the two strings. The normal value is 0, indicating no similarity and 1, indicating exact similarities (Leonardo and Hansun, 2017). The Jaro measurement is defined as follows (Friendly, 2019) :

$$Dj = \frac{1}{3}\left(\frac{c}{|s1|} + \frac{c}{|s2|} + \frac{c-t}{c}\right)$$

where,

c = Number of identical characters at the same position
$|s1|$ = The length of the first word
$|s2|$ = The length of the second word
T = ½ of the number of transpositions

Readopted consecutively for the development of a single terminology analysis algorithm.

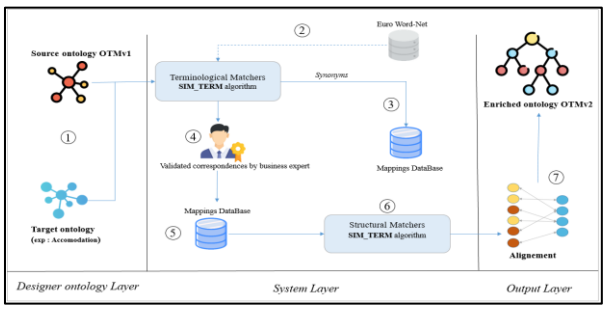

Fig. 3: The system architecture for aligning the OTMv1 ontology with the new business sub-ontologies identified

Algorithm 1: Algorithm SIM_TERM

Data:

S_O1*: The source ontology.*
S_O2*: The target ontology.*
F_RESEARCH_SYN*: A function returning a list of (synonyms, hypernyms and hyponyms) for a given concept.*
F_SIM_TERM*: A function that calculates terminological similarity by using the* ***distance of Jaro Winkler****.*
V_sc*: Vector of the terminological links for a given concept.*
S_st*: Structure of datum containing the information of similarity between two concepts.*
Tab_st*: List of terminological similarities.*

Result:

Tab_v_s1**:** *List of terminological similarities validated by the jobs experts.*

Begin:

```
/* Browse all concepts belonging to sub-ontology Source S_O1 */
For_each (C1 ∈ S_O1) {

   /* Search (synonyms, hypernyms…) of the selected concept */
   V_sc = F_RESEARCH_SYN (C1)

   /* Browse all concepts belonging to the sub-ontology Target S_O2 */
   For_each (C2 ∈ S_O2) {
      For_each (S1 ∈ V_SC) {

         /* Verify the existence of a similarity already calculated and validated by an expert in database mappings */
         S_st = VERIFY_BASE_MAP (C1, S1)

         IF S_stIs_Empty THEN
            IF (S1.type == C2.type) THEN

         /* Calculate the terminological similarity */
         SIM_TERM = CALCULSIMTERM (S1, C2, FC_SIM_TERM)
      /* Add S1, C2 and SIM_TERM to Tab_st*/
      Add ((S1, C2, SIM_TERM), Tab_st)
ELSE
   /* Add the similarity structure to Tab_st*/
   Add (S_st, Tab_st)
         }
         }
         }
/* The Results including into Tab_st must be validated by the jobs experts */
Tab_v_st = VALIDATION_RESULTS(Tab_st)

/* Inject the validated results into database mappings*/
INJECT_MAPPINGS (Tab_v_st)

END
```

The Alignment Algorithm Based on Terminological Matchers

The SIM_TERM algorithm calculates the terminological similarity of two concepts belonging to two ontologies, a source ontology and a second, so-called target ontology, related to tourism. The algorithm starts by searching for synsets of a given concept in the euro word-net lexical database (dictionary-based matcher), then calculates the terminological similarity between a concept in the source ontology and another in the target ontology based on the Jaro-Winkler distance (Jaro-Winkler distance-based matcher). Business experts then validate the results before being injected into the match database.

Structural Matchers

According to Univ Ctr of El Bayadh, Inst. Science and technology (Ardjani *et al*., 2015) these methods calculate the similarity between two entities by leveraging structural information when the involved entities are connected by semantic or syntactic links, creating a hierarchy or graph of entities we call:

- Internal structural methods: Methods that only exploit information on entity attributes
- External structural methods: Methods that take into account the relationships between entities

In this study, two external structural matchers are proposed:

Matcher Based on (Shvaiko *et al.*, 2007) *Study Results*

According to this study, two concepts can be considered similar if and only if:

Their super-concepts "father" are similar
Their sub-concepts "son" are similar
Their "neighboring" are similar

Matcher Based on Measurement 'Match-Based Similarity'

The function for calculating structural similarities is shown below (Tatane, 2023):

$$SIM\_STRUCT = \Sigma\, Pc_{(Gc,\, Gc')} XFC\_Sim\_Str(Gc,\, Gc')(Gc,\, Gc') \in (Tab\_vc,\, Tab\_vc')$$

where:

C : A concept belonging to source sub-ontology
C' : A concept belonging to the target sub-ontology
Tab_vc : Table of concepts (neighboring) for a concept *C* : Selected in the sub-ontologies S_01
Tab_vc': Table of concepts (neighboring) for a concept C' selected in the sub-ontologies S_02
Gc : Set of concepts (neighboring) belonging Tab_vc

Gc' : Set of concepts (neighboring) belonging Tab_vc'

Pc : Similarity weight of each category of concept.

$$FC\_Sim\_Str(Gc, Gc') = \Sigma\, Sim\_Term / Max\left(|Gc|, |Gc'|\right)(c, c') \in (Gc, Gc').\Sigma\, Pc_{(Gc, Gc')} = 1$$

N.B : *Gc and Ge' contains concepts (Neighboring) of the same category*

Alignment Algorithm Based on Structural Matchers

The SIM_STR algorithm calculates structural similarities between two domain sub-ontologies in graph mode. It starts by extracting the nodes of each sub-ontology, then looks for the neighboring nodes of each concept of the first sub-ontology and the neighbors of the concepts of the same category of the first concept matcher based on (Shvaiko *et al*., 2007) study results and calculates the structural similarity between these neighboring nodes (matcher 'match-based similarity').

Algorithm 2: Algorithm SIM_STR

Data:

***S_O1_G**: The source sub-ontology, graph format.*
***S_O2_G**: The target sub-ontology, graph format.*

***Pc**: similarity weight of each concept category.*

***Tab_c_oc**: List of concepts containing the source sub-ontology **S_O1**.*
***Tab_c_or**: List of concepts containing the target sub-ontology **S_O2**.*
***Tab_vc1**: List of Neighboring concepts for selected concept into **S_O1**.*
***Tab_vc2**: List of Neighboring concepts for selected concept into **S_O2**.*

***CALCUL_SIM_STRUCT**: A function to calculate structural similarities (Match-Based similarity).*

Result:
***Tab_ss*: List of structural similarities.**

Begin:

***S_O1_G** = **TRANSFORM_ONTOLOGY_GRAPH** (**S_O1**).*
***S_O2_G** = **TRANSFORM_ONTOLOGY_GRAPH** (**S_O2**).*

/ Extraction of concepts (nodes) contained in S_O1 */*
***Tab_c_oc** = SEARCH_CONCEPTS (**S_O1_G**).*
/ Extraction of concepts (nodes) contained in S_O2 */*
***Tab_c_or** = SEARCH_CONCEPTS (**S_O2_G**).*
***For_each** (concept∈**Tab_c_oc**) {*

/ Search Neighboring concepts starting with a top node of S_O1 */*
***Tab_vc1 = SEARCH_VOISINS** (Tab_c_oc[i])*

***For_each** (concept∈**Tab_c_or**) {*
IF** Tab_c_oc[i].type == Tab_c_or[i].type **THEN
/ Search Neighboring concepts starting with a top node of S_O2 */*
***Tab_vc2 = SEARCH_VOISINS** (Tab_c_or[i])*

/ Calculate structural similarities */*
Sim_struct = CALCUL_SIM_STRUCT (**Tab_vc1 [i], Tab_vc2 [i], Pc**)

/ Add concept1, concept2 and the value of sim_struct to Tab_ss*/*
Add ((**Tab_vc1 [i], Tab_vc2 [i], Sim_struct), Tab_ss**)
}
}

Return (Tab_ss)

END

Materials and Methods

Otmv1

Ontology of tourism in Morocco was created through manual analysis of the semantics of a specialized thesaurus. This was part of a research project called knowledge management and Web Semantic-GECO-WES [1], launched by the IRF-SIC laboratory at the faculty of science, Ibn Zohr University, Morocco.

Tree tagger

The tree-tagger tool is a program that can be used on windows or UNIX via the command prompt. We used the UNIX version (http://www.ims.uni-stuttgart.de/).

YaTeA

YaTea is a term extractor that identifies and extracts noun phrases that could be term candidates. Each term is syntactically analyzed to reveal its structure in the form of heads and modifiers (http://perso.limsi.fr/hamon/YaTeA/).

TermoStat

Termostat is a multi-language term extractor available on the internet. It is based on linguistic knowledge and compares the use of a term in a specialized corpus with its presence in a general language corpus to determine its relevance (http://termostat.ling.umontreal.ca/).

Gate-Developer

The functionalities provided by the gate-developer text-engineering platform and its 'JAPE' workshop was used to define a set of lexico-syntactic patterns that have automated the process of identifying relations between ontological concepts (http://www.gate.ac.uk/download/).

WOLF Lexical Database

Free French lexical database was used to explore the relationships between terms in the corpus.

The FACT ++ Reasoner of Protégé 2000

FACT++ reasoner included in the protégé 2000 tool was used for the analysis of the conformity and coherence of the new ontological model studied.

Results

The experimentation of the proposed approach has resulted in upgrading the initial ontological core and developing a new version (OTMv2), enriched with increased conceptual diversity and depth. The need to validate and approve this new model led us to the conception and creation of a study platform, which we named (OTMv2 WebApp), offering experts (remotely connected) the ability to access, visualize, query and align the base ontology with other (ontologies/sub-ontologies) identified during the enrichment phase. The platform was initially launched online at http://38.242.226.142/otm/. It will soon be suspended to optimize its performance and ensure the confidentiality of the research and development work currently being conducted by our research team.

It is worth noting that the technical tools used for the development of OTMv2 WebApp include: JEE, eclipse, tomcat, maven, SVN, MySQL, MariaDB, spring MVC, spring security and hibernate. The ontology upgrade process proposed for the OTMv2 WebApp platform begins with (remote) authentication of the business expert and consists, as shown in Fig. 4, of the following three main phases.

Ontology Loading

The aim of this module is to load the OTMv1 base ontology and the search ontology (source file in OWL or XML format) and present them to business experts (model validation) in various usable forms (tables, graphs, etc.,) thus facilitating their reading, analysis, verification and validation.

In this context, a set of functionalities useful for the analysis and understanding of the ontological base and search models have been developed, Fig. 5, namely: Graphical visualization of the ontology, separation of ontological components (classes, individuals, data and object properties) and execution of queries on ontological models, both the base and search models.

Ontology Analysis

The aim of this module is to facilitate understanding and analysis of the ontological models proposed for an alignment study, the OTMv1 base ontology and its business sub-ontologies. It enables visualization, as shown in Figs. 6-7 and semantic interrogation of these models under study.

The OTMv2 WebApp platform provides business experts with two reading modes for analyzing ontological models under study.

Direct interpretation of the loaded OWL or XML file, Fig. 6.

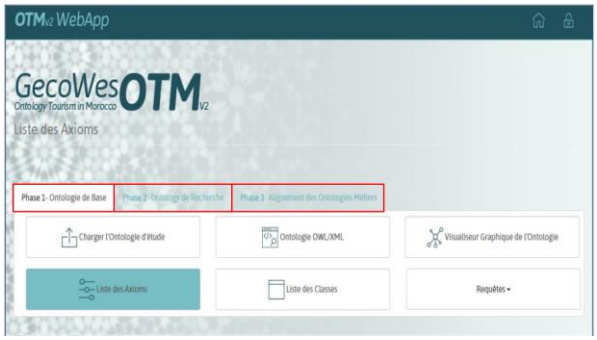

Fig. 4: Extract from the main modules of the OTMv2 WebApp platform

Fig. 5: Main functionalities offered to the business experts for reading and analyzing the basic and search ontologies

Exploration of the hierarchical structure, based on the ontological components presented in the loaded source file, as shown in Fig. 7. Semantic querying is used to analyze the semantic capabilities of the ontological models in question, using two query languages:

- DLquery (concerns global semantic entities), as shown in Fig. 8
- SPARQL (for semantic instances) Fig. 9

In the example illustrated in Fig. 8, the business expert is looking to examine the semantic capabilities of the concept

of tourist accommodation. In this context, he has formulated the following semantic question: "Which rural accommodation establishments allow the presence of pets?".

The answers provided were based on semantics, with certain restrictions linked to the target concept, i.e., "accommodation (Rural, pet-friendly)", as shown in the results section.

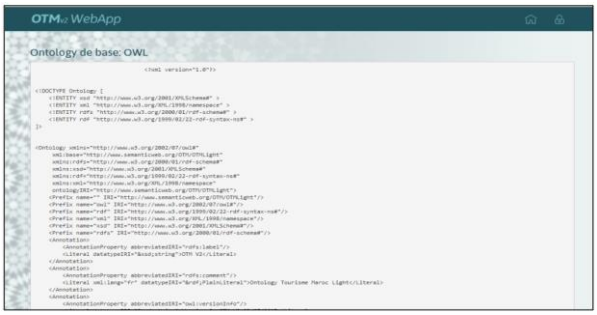

Fig. 6: Basic view mode, read OWL or XML source file

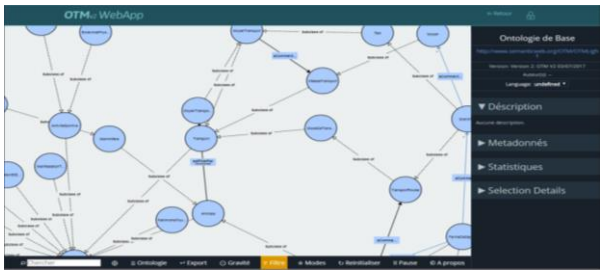

Fig. 7: Graphical display of the loaded ontology

Fig. 8: Example of a DLquery request: “Which rural accommodations accept pets

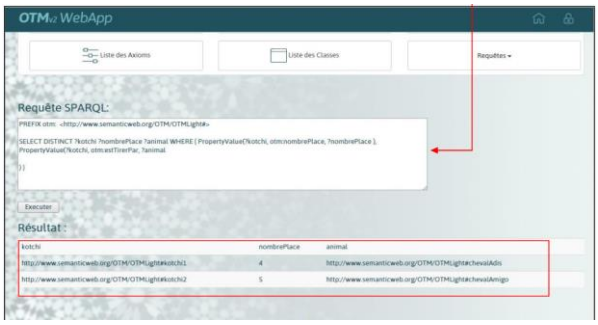

Fig. 9: Example of the execution of a SPARQL query what are the means of urban transport pulled by an animal

Fig. 10: The main steps proposed for aligning business sub-ontologies

In the example shown in Fig. 9, the business expert is seeking to analyze the semantic capabilities of the concept "transport". In this context, he has formulated the semantic query under the SPARQL query language as follows: "Animal-pulled means of urban transport?"

Exploration of the ontological model generated extremely precise semantic responses, referring to information such as the registration number of the Kotchi (a means of transport specific to the Moroccan context), the number of seats for each instance and the name of the horse responsible for the transfer operation, as shown in the results section.

Aligning Evolving Business Sub-Ontologies

The aim of this module is to implement the methodological approach to alignment proposed in this study. As shown in Fig. 10, the process begins by loading the ontological entities of the base model and search model into the so-called mapping database, followed by the launch of terminology matchers. The results of the third stage are then validated by the business experts, as shown in Fig. 11 and finally, the structural matchers are launched, as shown in Fig. 12. Finally, the concepts of the ontological models under study are dynamically mapped, as shown in Fig. 13. The OTMv2 WebApp application offers five processing steps:

- Loading ontological concepts from the base and search ontologies into the temporary mapping database
- Executing terminology matchers
- Validation of terminology results by the business expert, as shown in Fig. 11
- Structural matchers are launched based on the previously validated results, as shown in Fig. 12
- The dynamic creation of correspondences between concepts in the ontological models under study

Following the sequential execution of the terminology functions mentioned above, the results presented in raw form are submitted to the business expert for further clarification, accompanied by preliminary alignment suggestions, as shown in Fig. 11. The validated conceptual pairs are saved in the final match database

before being used as input for the structural analysis functions, as shown in Fig. 12.

Semantic links with significant structural similarity values are dynamically declared and then established within the overall ontological model.

Despite the advantages offered by this platform, we emphasize that several areas for improvement could be subject to research and development, such as The automated detection and identification of ontologies already developed and validated by other projects and thus exploring the possibility of creating dynamic mapping based on the history of mappings already handled by the domain experts consulted.

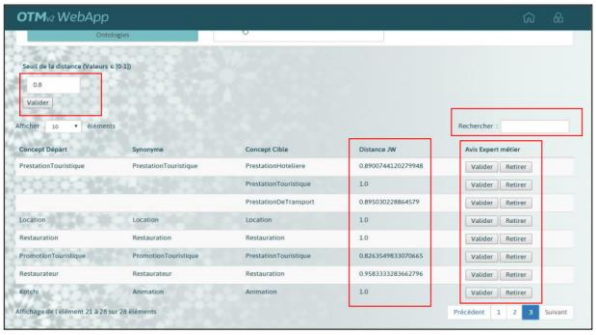

Fig. 11: Example of results obtained after applying terminological matchers, pending validation by business experts

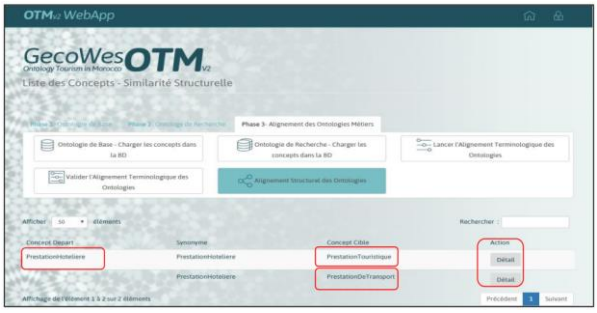

Fig. 12: Example of terminology alignment results validated by business experts, proposed for structural alignment

Fig. 13: Example of mapping results between two concepts "hotel service" and "tourist service" belonging to two distant ontologies

Discussion

As with any research project, we believe that certain perspectives are brought to light. In continuation of this study, we deem it important to pursue the study of the following avenues:

- Exploiting data invalidated by domain experts during the terminological analysis phase. We posit that highlighting rejected elements in this context can expedite the treatments and future interventions of the same type
- Examining the quality of a hybrid ontological version constructed entirely based on external business ontologies already developed and validated by other research communities
- The heterogeneity and distributed nature of local ontology modeling in practice necessitate a personalized parameterization of the variables proposed by our system. We believe that the flexibility to adjust study parameters can be highly beneficial. We hold that the thresholds defined for similarity calculation can be modified based on the domain of application and/or the specificity of the modeling and anticipated results
- Analyzing the impact of the evolution of local ontologies on previously validated associations as well as on the consensus ontology. In this context, we believe that the excessive evolution of sub-ontologies may lead to the creation of vertical mappings between a local ontology and the global ontology

Conclusion

Enriching and populating ontologies represent a developing research area. Therefore, we sought to establish a new methodological approach to update a core ontology covering Moroccan tourism and its specificities. The work proposed in this context is based on two phases of study.

The first phase involves analyzing and extracting new conceptual and semantic entities from a representative textual corpus of the domain, following a well-studied exploitation of a suite of natural language processing tools.

The second phase aims at aligning the base ontology with other identified business sub-ontologies following a thorough analysis of the conceptual structure of OTMV1. This led us to design and develop a series of specific algorithms to propose probable mappings (terminological and structural) to domain business experts.

Experimentation with this approach facilitated the extension of the initial ontological model through the creation of a new version that is more representative of the domain and semantically richer. However, the

development of the WebApp OTMV2 platform ensures remote involvement of domain experts and thus proposes a set of technical tools to facilitate understanding, analysis and interrogation of the ontological model under study, in addition to technical indicators to assist cognitive scientists during the mapping validation phase.

Acknowledgment

The authors wish to express their gratitude to the editors for their diligent efforts in managing the manuscript and to all the reviewers for their valuable feedback, which has significantly improved the original submission.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Khalid Tatane: Contributed to the research, design, implementation of the proposed algorithm, analysis of results and written of the manuscript.

Asma Amalki: Contributed to the research, design of the study, literature review, participated in all experiments, coordinated the data-analysis, result evaluation and written the manuscript.

Ali Bouzit: Data analysis reviewed and results evaluation, designed the research planed and organized the study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

References

Ardjani, F., Bouchiha, D., & Malki, M. (2015). Ontology-Alignment Techniques: Survey and Analysis. *International Journal of Modern Education and Computer Science*, *7*(11), 67-78. https://doi.org/10.5815/ijmecs.2015.11.08

Bailal, H., Boumeska, M., & Lahssini, M. (2023). L'étude de l'influence du système de contrôle de gestion sur la gouvernance des collectivités territoriales : Cas des régions du Maroc. *[RMd] RevistaMultidisciplinar*, *5*(3), 167-195. https://doi.org/10.23882/rmd.23169

Sree Harissh, M. Vignesh, U. Kodaikkaavirinaadan, T. V. Geetha (2018). Unsupervised Domain Ontology Learning from Text, *Polibits*, Vol. *57*, p. 59 66, janv. https://doi.org/10.17562/PB-57-6

Bellefleur, M. (2011). *Du Rapport Bélisle À La Création Du Haut-Commissariat À La Jeunesse, Au Loisir Et Au Sport* (1st Ed., p. 12). Presses de l'Université du Québec. https://doi.org/10.2307/j.ctv18phcj6.12

Cardellino, C., Teruel, M., Alemany, L. A., & Villata, S. (2017). Ontology Population and Alignment for the Legal Domain: YAGO, Wikipedia and LKIF. *ISWC*. Posters, Demos and amp; Industry Tracks. https://iswc2017.ai.wu.ac.at/wp-content/uploads/papers/PostersDemos/paper636.pdf

Chen, H., & Luo, X. (2019). An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. *Advanced Engineering Informatics*, *42*, 100959. https://doi.org/10.1016/j.aei.2019.100959

Clancy, B., & Vaughan, E. (2023). *Using Corpus Linguistics to Interpret Economic News Texts* (1st Ed, pp. 166-191). Routledge. https://doi.org/10.4324/9781003154747-10

Dai, G., & Liu, S. (2024). Towards Predicting Post-Editing Effort with Source Text Readability. *The Journal of Specialised Translation*, *41*, 206-229. https://doi.org/10.26034/cm.jostrans.2024.4723

Djellali, C. (2013). Using hamming similarity to map ontology learning: A new data mining system. *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, 82-87. https://doi.org/10.1145/2513228.2513232

Elbacha, M. (2023). Nouvelle Méthode d'Extraction Automatique Bilingue des Syntagmes Terminologiques Nominaux à Base de leurs Noyaux et du Balisage Structurel XML du Corpus Aligné. *The Egyptian Journal of Language Engineering*, *10*(2), 51-68. https://doi.org/10.21608/ejle.2023.234311.1054

Essayeh, A., & Abed, M. (2015). Towards Ontology Matching Based System Through Terminological, Structural and Semantic Level. *Procedia Computer Science*, *60*, 403-412. https://doi.org/10.1016/j.procs.2015.08.154

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association*, *23*(5), 1007-1015. https://doi.org/10.1093/jamia/ocv180

Friendly, F. (2019). Jaro-Winkler Distance Improvement for Approximate String Search Using Indexing Data for Multiuser Application. *Journal of Physics: Conference Series*, *1361*(1), 012080. https://doi.org/10.1088/1742-6596/1361/1/012080

Gao, S., & Liu, Y. (2023). A people-item relation extraction method based on multiple kernel support vector machine model. *5th International Conference on Artificial Intelligence and Computer Science (AICS 2023)*, 1280319. https://doi.org/10.1117/12.3009553

Gašić, M., Hakkani-Tür, D., & Celikyilmaz, A. (2017). Spoken language understanding and interaction: Machine learning for human-like conversational systems. *Computer Speech and amp; Language*, *46*, 249-251. https://doi.org/10.1016/j.csl.2017.05.006

Labidi, N., Chaari, T., & Bouaziz, R. (2017). An NLP-Based Ontology Population for Intentional Structure. In A. M. Madureira, A. Abraham, D. Gamboa, & P. Novais (Eds.), *Intelligent Systems Design and Applications* (1st Ed, Vol. *557*, pp. 900-910). Springer International Publishing. https://doi.org/10.1007/978-3-319-53480-0_89

Leonardo, B., & Hansun, S. (2017). Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, *5*(2), 462. https://doi.org/10.11591/ijeecs.v5.i2.pp462-471

Li, P., & Mao, K. (2019). Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, *115*, 512-523. https://doi.org/10.1016/j.eswa.2018.08.009

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). Biomedical term extraction: Overview and a new methodology. *Information Retrieval Journal*, *19*(1-2), 59-99. https://doi.org/10.1007/s10791-015-9262-2

Makki, J. (2017). Ontoprima: A Prototype for Automating Ontology Population. *International Journal of Web and amp; Semantic Technology*, *8*(4), 1-11. https://doi.org/10.5121/ijwest.2017.8401

Moens, M. F. (2018). Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument and AMP Computation*, *9*(1), 1-14. https://doi.org/10.3233/aac-170025

Nasr, S. (2023). L'extraction terminologique automatique : Une approche centrée sur l'apprenant. *The Egyptian Journal of Language Engineering*, *10*(1), 10-35. https://doi.org/10.21608/ejle.2023.175587.1037

Ngo, D., & Bellahsene, Z. (2016). Overview of YAM++—(not) Yet Another Matcher for ontology alignment task. *Journal of Web Semantics*, *41*, 30-49. https://doi.org/10.1016/j.websem.2016.09.002

Portisch, J., Hladik, M., & Paulheim, H. (2022). Background knowledge in ontology matching: A survey. *Semantic Web*, *15*, 1-55. https://doi.org/10.3233/sw-223085

Pressat-Laffouilhère, T. (2023). *Modèle ontologique formel, un appui à la sélection des variables pour la construction des modèles multivariés*. https://theses.hal.science/tel-04500818

Shvaiko, P., Euzenat, J., Stuckenschmidt, H., Mochol, M., Giunchiglia, F., Avesani, P., Van Hage, W. R., Sváb, O., Svátek, V., & Yatskevich, M. (2007). Description of alignment evaluation and benchmarking results. *HAL Open Science*, 69. https://inria.hal.science/hal-00822894/document

Tatane, K., Amalki, A., & Bouzit, A. (2023). L'impact de la prise en charge des couches sémantiques sur la recherche d'information touristique: Cas des établissements et services touristiques digitalisés au Maroc. *[RMd] RevistaMultidisciplinar*, *5*(3), 259-289. https://doi.org/10.23882/rmd.23172

TWBD. (2023). *The World Bank DATA*. https://data.worldbank.org/topic/tourism

Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers and amp; Geosciences*, *112*, 112-120. https://doi.org/10.1016/j.cageo.2017.12.007

Wang, Y., Qin, J., & Wang, W. (2017). Efficient Approximate Entity Matching Using Jaro-Winkler Distance. In A. Bouguettaya, Y. Gao, A. Klimenko, L. Chen, X. Zhang, F. Dzerzhinskiy, W. Jia, S. V. Klimenko, & Q. Li (Eds.), *Web Information Systems Engineering-Wise 2017* (1st Ed, Vol. *10569*, 231-239). Springer International Publishing. https://doi.org/10.1007/978-3-319-68783-4_16

Zhou, M. (2018). What Will Search Engines be Changed by NLP Advancements. *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, 7. https://doi.org/10.1145/3234944.3241521

Zhu, Y., Li, X., Wang, Z., Li, J., Yan, C., & Zhang, Y. (2023). ER-LAC: Span-Based Joint Entity and Relation Extraction Model with Multi-Level Lexical and Attention on Context Features. *Applied Sciences*, *13*(18), 10538. https://doi.org/10.3390/app131810538