INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH AND INNOVATIVE STUDIES
ISSN: 2820-7157
www.ijsrisjournal.com

June 2024 **Volume 3** **Number 3** **Pages 41-45**

Received Date: April 11, 2024 **Accepted Date:** May 10, 2024 **Published Date:** June 01, 2024

Available Online at https://www.ijsrisjournal.com/index.php/ojsfiles/article/view/155

DOI: https://doi.org/10.5281/zenodo.11244423

Semantic Enrichment of Moroccan Tourism Ontology Based NLP Technologies

ASMA AMALKI[1], KHALID TATANE[2], ALI BOUZIT[3]

[1]Image and Pattern Recognition – Intelligent and Communicating Systems Laboratory (IRF-SIC), Faculty of Science, Ibn Zohr University, Agadir, Morocco, asma.amalki@edu.uiz.ac.ma

[2]National School of Applied Sciences, ESTIDMA research team, Ibn Zohr University, Agadir, Morocco, k.tatane@uiz.ac.ma

[3]Image and Pattern Recognition – Intelligent and Communicating Systems Laboratory (IRF-SIC), Faculty of Science, Ibn Zohr University, Agadir, Morocco, a.bouzit@uiz.ac.ma

ABSTRACT

Ontology is a conceptual representation model that allows sharing and reuse of a domain knowledge in a human and machine-readable format. However, the massive amount of knowledge available today makes ontology enrichment a challenging task. In this paper, we present a semi-automatic approach of ontology learning from collection of domain specific texts, to text processing based NLP tools, to enrich an existing ontology. This study utilize data crawling from official websites. Concepts and relations extraction is done automatically from a textual corpus and domain experts do consistency and redundancy check manually. The proposed approach is applied in order to enrich Moroccan tourism ontology named OTM with semantic entities extracted from heterogeneous data sources.

Keywords: Corpus, Knowledge acquisition, NLP, ontology enrichment, ontology learning.

1. INTRODUCTION

Currently, tourism information is widely dispersed. Searching for this information typically involves spending a considerable amount of time sorting through search engine results, selecting, and reviewing the details of each accommodation [1]. In this context, employing ontology representation can facilitate user searches, allowing for more precise search results. Ontologies and associated semantic web technologies play a vital role in facilitating the conversion of data from human-readable to machine-readable formats. They empower sophisticated information processing and management techniques [2]. Ontology is a conceptual representation; it is a set of relations and concepts describing a specified domain. It is formally defined, rendering it accessible to machines; it is widely accepted by the scientific community [2]. Ontology aims to ensure the coherence and clarity of knowledge and offers a framework for both sharing and integrating knowledge [3].

At present, there is a rapid advancement in information technology within the tourism industry, every day a massive data is generated by humans and automated agents in various formats. Therefore, tourism ontologies should be enriched regularly. Knowledge managers and domain experts typically build ontology. However, the process of identifying concepts, establishing relationships, and incorporating new instances into the ontology is arduous, time-consuming, tedious, cost-intensive, and intricate. Consequently, there is an increasing interest in (semi)automated methods for learning and enriching ontologies [4]. The main challenge in this process is that knowledge is highly unstructured and difficult to convert in meaningful model.

Ontology enrichment involves incorporating new instances of concepts into an ontology without altering its structure. As a result, neither the conceptual hierarchy nor the non-taxonomic

relationships are modified after the population process. This process necessitates an existing ontology and an extraction engine that analyzes data to identify objects linked with concepts. Various methodologies have been proposed within the realm of ontology learning and evolution, reflecting its critical applications across diverse domains and making it a highly active area of research [2].

In this paper, we present a semi-automatic approach for ontology enrichment based on terminological, relational, and semantic analysis of a study corpus related to Moroccan tourism using Natural Language Processing tools. This research work was carried out by [5].

The remainder of this article is structured as follows: Part 2 presents the state of the art, Part 3 describe the proposed work, and Part 4 includes the results of the proposed approach. In the last part, we present a general conclusion outlines our future interventions in this field.

2. STATE OF THE ART

2.1 Ontology of Tourism in Morocco

OTMv1 (Ontology of Tourism in Morocco version1) was developed from scratch by analyzing the semantics of a specialized thesaurus created and developed by The World Tourism Organization as part of a research project entitled Knowledge Management and Web Semantic - GECO-WES [6], initiated by the laboratory (IRF-SIC) of the Faculty of Science at Ibn Zohr University. The study presented in this paper proposed by [5] has allowed a generation of a new version of OTMv1 kernel more expressive and semantically richer through the semantic analysis of a textual corpus.

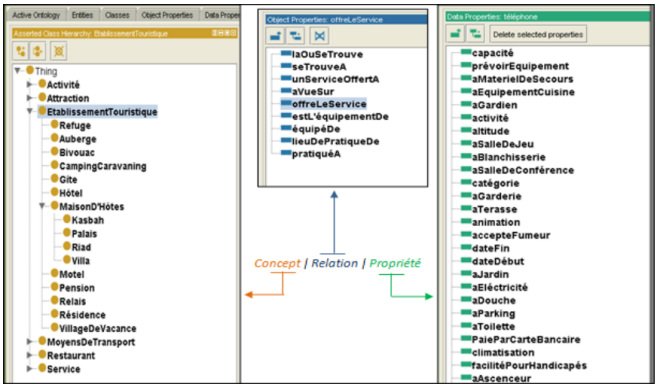

Figure 1 : Ontology of Tourism in Morocco version1 [6]

2.2 Knowledge Acquisition

According to [7] Knowledge acquisition from raw input resources is required by ontology enrichment, performing this process manually is both costly and prone to errors. Information extraction (IE) offers automation for knowledge acquisition. IE systems analyze a diverse corpus as input, extracting instances of concepts and relationships from each knowledge resource within the corpus, including text, images, videos, and audio files. Consequently, various IE systems are employed to extract information from diverse documents, after identifying the knowledge source and employing the relevant extractor. The information extracted by an IE system can then be utilized to enrich the ontology with new instances of concepts and relationships. Methods for extracting domain-specific terms, concepts, and their associations can be categorize into two types: linguistic techniques, statistical techniques and the Hybrid or mixed method [8]. The first type operates under the assumption that grammatical structure reflects semantic dependencies, intending to identify the semantic dependencies among terms in a sentence. These methods analyze the grammatical function of words within a sentence, enabling them to extract terms and their relationships [9]. The second type identify terms based on their distribution in the texts. These statistical techniques are derived from data mining, machine learning, and information retrieval. While these methods can identify new candidate terms for ontology enrichment, they are unable to incorporate them into the ontology without human intervention [9]. The last one is generated as a natural result of the fusion of the two previous methods, to fill in the gaps and exploit the advantages [8].

2.3 Ontology Learning

Ontology Learning is an approach aimed at rapidly and effectively constructing ontologies by utilizing automated techniques to convert domain knowledge into ontology structures. It encompasses several components including Ontology Enrichment, Ontology Population, Ontology Alignment and Ontology Merging[10]. Ontology enrichment involves expanding an established ontology by incorporating supplementary concepts and semantic relationships, positioning them appropriately within the ontology. On the other hand, ontology population entails the addition of new instances of concepts to the ontology [4]. Significant research attention has been devoted to ontology enrichment and population, given their crucial roles in information extraction, semantic annotation, and search [4]. The ontology enrichment process can be divided into three stages: a first phase of knowledge acquisition or ontology learning, then a second phase of candidate term validation, and finally a third phase of integration into the ontology [11].

Ontology alignment resolves the issue of semantic heterogeneity by establishing semantic connections between entities of different ontologies. Given two ontologies O1 and O2, the process of ontology alignment is to find mappings between concepts from O1 and O2. A mapping can be described with a four tuple, as written in (1), where c1 and c2 are concepts of O1 and O2, respectively [12].

$$m = \langle id, c1, c2, s \rangle \quad (1)$$

Where:

id indicates the unique identifier of each mapping

$s \in [0, 1]$ is the degree of confidence of the alignment relation (equivalence or subsumption) between *c1* and *c2*.

Ontology Merging involves consolidating compatible concepts into a unified concept, thereby generating a single ontology from two separate ones. The semantic matches described in the designations may pertain to equivalence (is-a), specialization, and/or generalization (part of) relationships, among other senses [13].

2.4 Ontology Enrichment Based NLP

In recent years, Natural Language Processing (NLP) has been extensively explored as an effective technology for text mining, facilitating knowledge extraction from unstructured text [14]. Natural Language Processing (NLP) has been employed and experimented with to acquire knowledge from textual data, particularly in the domain of ontology learning, it have demonstrated encouraging results [15]. In general, NLP methods can be classified into:

Symbolic methods: employs linguistic information for extracting information from text. For example, noun phrases are viewed as lexicalized concepts and frequently employed to represent concepts in an ontology. Linguistic rules that describe the relations between terms in the text can also aid in recognizing conceptual relationships within an ontology [15]. The most used symbolic approaches are lexico-syntactic patterns that indicate superficial relational markers present in a given natural language (for example :"such as") and internal syntactic structure of component terms that can indicates hyponymy relation between terms [15].

Statistical methods: relies on large textual corpus, it employs various linguistic principles and features to statistically measure and extract semantic information. One such linguistic principle is selectional restrictions, where syntactic structures offer pertinent insights into semantic content [15]. Statistical methods can be categorized into clustering approaches and machine learning methods. Clustering approaches use similarity measurements that can offer insights into hierarchical and synonyms relationships of concepts. Machine learning approaches treat the knowledge extraction problem as a classification process [15].

Hybrid methods: combines symbolic and statistical methods.

Several NLP methods have been utilized for ontology enrichment in many studies, [16] is a study that presents an automatic approach NLP-based for enriching an ontology of intentions from textual client request in IT market. This approach have used chunking to extract semantic components from informal text, POS (Part-of-Speech) tagging to assign grammatical categories to every term in extracted sentences and linguistic rules to classify extracted terms in its suitable concepts in the intention ontology. ONTOPRIMA [14] is a semi-automatic approach for ontology enrichment in the domain of risk management, it have use NLP techniques for language processing applied on textual corpus built from chemical facts sheets issued by the U.S. Environmental Protection Agency. [2] is an automatic and multi-modality approach proposed to enrich an ontology of multimedia domain named ImageNet. It uses NLP techniques for combining textual and visual information and convolutional neural network for features extraction task.

3. NLP-BASED ONTOLOGY ENRICHMENT FOR MOROCCAN TOURISM ONTOLOGY

The proposed semi-automatic enrichment method for the Moroccan tourism ontology OTMv1 aims to expand its foundational core by incorporating additional concepts and semantic relationships derived from unstructured texts. Figure 2 illustrates the six stages comprising this approach.

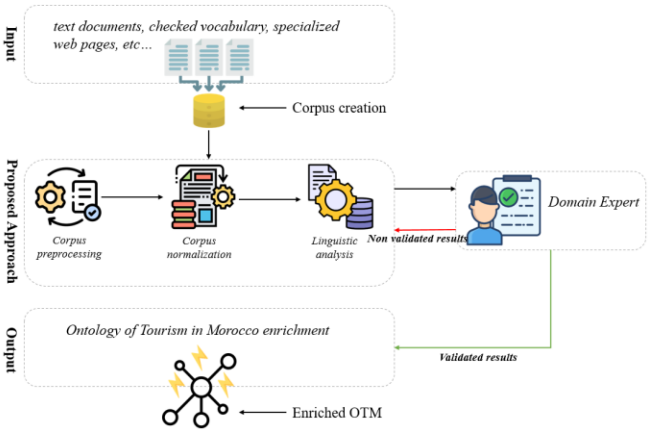

Figure 2 : proposed approach for OTMv1 enrichment based NLP tools

3.1 Creation of Specialized Corpus

The corpus was created from diverse and specialized sources such as blogs, web portals, text files, electronic newspapers, and magazines, among others by web crawling. This broad selection was undertaken to encompass a wide range of knowledge pertaining to Moroccan tourism. Summary of utilized knowledge sources: 300 web pages, 10 PDF files, 8 PPT files, and 10 tourism magazines. Gathered from the official websites of the following organizations:

- Ministry of Tourism, Handicrafts, and the Social and Solidarity Economy of Morocco.
- High Commission for Planning of the Kingdom of Morocco.
- UNWTO World Tourism Barometer.
- World Bank - Tourism data.

The collected data was over a million words to analyze.

3.2 Corpus Pre-processing

The unstructured text collected had to go through a number of pre-processing stages to ensure a proper semantic analysis of the text. These steps can be indicated as follows:

- Unification of HTML, PDF and PPT formats into a single, native text format.
- 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.

This stage of corpus pre-processing reduce lexical inconsistencies and syntactic ambiguities from the initial text.

3.3 Corpus Standardization

Standardization of the heterogeneous corpus is a key step in the automatic language processing process. Two essential operations was used for this purpose:

- Word lemmatization: Lemmatization involves reducing a word to its canonical form, typically, as it appears in a dictionary. This process helps reduce morphological variations within text, making tasks like semantic analysis, information retrieval, question answering, or search operations more efficient [17].
- Sentence segmentation: sentence segmentation is a key operation in the NLP pipeline; it is defined as the process of dividing a chunk of text into meaningful sections according to their topical continuity. This task constitutes a fundamental challenge in natural language processing (NLP) and is applied in various tasks such as summarization, passage extraction, discourse analysis, question answering, context comprehension, and document noise reduction [18].

Tree Tagger has been used to implement the two operations mentioned above. This stage was led to reduction of lexical inconsistencies and syntactic ambiguities from the corpus.

3.4 Linguistic Analysis

Linguistic analysis procedure have to go from those stages:

- Morph syntactic tagging of texts using the Tree Tagger tool.
- Extraction of candidate terms using the term extractors TermoStat.
- Extraction of conceptual relationships (synonyms and hyperonymy relationships) based on the lexical database WOLF.

3.5 Semantic Standardization

This phase requires the intervention of the domain expert to:

- Validate the candidate terms extracted in the previous stage.
- Validate the lexical relationships extracted in the previous stage.

After validation of candidate terms and lexical relationships by the domain expert, the validated elements are processed as follows:

- Validated candidates terms are converted to concepts
- Validated lexical relationships are converted to semantic conceptual relations.

3.6 OTMv1 Core Enrichment

The new extracted concepts and semantic relationships were integrated into OTMv1 ontology using OWL language and knowledge processing software Protégé 2000. FACT++ reasoner of Protégé 2000 was used for Compliance and consistency analysis of the new enriched ontological model.

4. Results

The methodological approach introduced in this paper facilitated the semi-automatic enhancement of the initial OTMv1 ontology developed within the GECO-WES project by introducing additional concepts and relationships, as illustrated in Figure 3. The procedure commenced with the construction of a comprehensive tourism corpus, comprising more than a million entities for analysis. Employing NLP tools on this corpus allowed the extraction of 4412 potential terms and 1114 lexical relationships. The candidate terms relevant to the Moroccan tourism context were transformed into concepts, and the lexical relationships were converted into semantic relations. Following validation of the results, along with evaluation and integration of the new semantic entities into the foundational ontological model, the entire process effectively expanded the core OTMv1 ontology, enhancing its expressiveness and semantic depth.

Figure 3 : New concepts identified and added to OTMv1

Utilizing the proposed methodology, the newly developed OTMv2 ontology model exhibits enhanced semantic depth. It incorporates more than twenty novel generic business concepts, thereby expanding the semantic coverage for future inquiries within the realm of national tourism. In this work, the reasoning was done by using fact++ plugin in Protégé, the results prove that OTMv2 was consistent.

5. Conclusion

Enriching and populating ontologies is an evolving area of research. Thus, our aim was to present a methodological approach for enhancing a fundamental ontology that addresses Moroccan tourism and its particular characteristics. In this context, a specialized corpus was built from heterogeneous and unstructured text, then preprocessed and normalized to

increase its quality. Linguistic analysis was applied on the preprocessed corpus for semantic entities extraction based on NLP methods and tools. The extracted terms and relationships was converted then to concepts and semantic relationships under the supervision of a domain expert. The new extracted concepts and relationships was then integrated to the OTMv1 kernel. The new version of the OTM ontology model, created through the proposed methodology, boasts enhanced semantic depth. It incorporates more than twenty novel generic business concepts, providing a wider semantic range for forthcoming inquiries within the realm of national tourism.

The ongoing enhancement of the Moroccan tourism ontology, considering the growth of accessible knowledge and the essential intervention from domain experts, poses a significant challenge in terms of time and resources required. Therefore, there is a demand for a methodological approach that automates the process of enriching the ontology core entirely.

REFERENCES

1. Q. Li, S. Li, S. Zhang, J. Hu, et J. Hu, « **A Review of Text Corpus-Based Tourism Big Data Mining** », *Applied Sciences*, vol. 9, no 16, p. 3300, août 2019, doi: 10.3390/app9163300.
2. M. Muscetti, A. M. Rinaldi, C. Russo, et C. Tommasino, « **Multimedia ontology population through semantic analysis and hierarchical deep features extraction techniques** », *Knowl Inf Syst*, vol. 64, no 5, p. 1283-1303, mai 2022, doi: 10.1007/s10115-022-01669-6.
3. *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*: proceedings : October 19-23, 2020, Bandung - Padang, Indonesia. Piscataway, NJ: IEEE, 2020.
4. M. Kokla, V. Papadias, et E. Tomai, « **ENRICHMENT AND POPULATION OF A GEOSPATIAL ONTOLOGY FOR SEMANTIC INFORMATION EXTRACTION** », Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLII-4, p. 309-314, sept. 2018, doi: 10.5194/isprs-archives-XLII-4-309-2018.
5. K. TATANE, « **Amélioration du processus d'analyse sémantique à travers le développement incrémental d'une ontologie du domaine basé sur des approches d'enrichissement et d'alignement : Application au domaine touristique marocain** », Thèse, IBN ZOHR, AGADIR, MAROC, 2017.
6. S. MOUHIM, « **Une approche de la gestion de connaissances basée sur les ontologies, les technologies du web sémantique et les démarches linguistiques : Application au domaine du tourisme.** », IBN ZOHR, AGADIR, MAROC, 2014.
7. M. Lubani, S. A. M. Noah, et R. Mahmud, « **Ontology population: Approaches and design aspects** », *Journal of Information Science*, vol. 45, no 4, p. 502-515, août 2019, doi: 10.1177/0165551518801819.
8. M. Elbacha, « **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*, vol. 10, no 2, p. 51-68, oct. 2023, doi: 10.21608/ejle.2023.234311.1054.
9. A. Ayadi, A. Samet, F. D. B. De Beuvron, et C. Zanni-Merk, « **Ontology population with deep learning-based NLP: a case study on the Biomolecular Network Ontology** », *Procedia Computer Science*, vol. 159, p. 572-581, 2019, doi: 10.1016/j.procs.2019.09.212.
10. *ICICoS 2019: the 3rd International Conference on Informatics and Computational Sciences* : proceedings : October 29th -30th, 2019, Semarang, Central Java, Indonesia. Piscataway, NJ: IEEE, 2019.
11. A. Tissaoui et A. Mezni, « **Ontological and Terminological Ressource Enrichment from Text Copora** », in 2016 *Global Summit on Computer & Information Technology (GSCIT)*, Sousse, Tunisia: IEEE, juill. 2016, p. 21-26. doi: 10.1109/GSCIT.2016.18.
12. Z. Hao, W. Mayer, J. Xia, G. Li, L. Qin, et Z. Feng, « **Ontology alignment with semantic and structural embeddings** », *Journal of Web Semantics*, vol. 78, p. 100798, oct. 2023, doi: 10.1016/j.websem.2023.100798.
13. M. Maroun, « **A survey on ontology operations techniques** », *Mathematical and Software Engineering*, vol. 7, no 1-2, p. 7-28, 2021.
14. J. Makki, « **ONTOPRIMA: A Prototype for Automating Ontology Population** », *IJWesT*, vol. 8, no 4, p. 1-11, oct. 2017, doi: 10.5121/ijwest.2017.8401.
15. K. Liu, W. R. Hogan, et R. S. Crowley, « **Natural Language Processing methods and systems for biomedical ontology learning** », *Journal of Biomedical Informatics*, vol. 44, no 1, p. 163-179, févr. 2011, doi: 10.1016/j.jbi.2010.07.006.
16. N. Labidi, T. Chaari, et R. Bouaziz, « **An NLP-Based Ontology Population for Intentional Structure** », *in Intelligent Systems Design and Applications*, vol. 557, A. M. Madureira, A. Abraham, D. Gamboa, et P. Novais, Éd., in Advances in Intelligent Systems and Computing, vol. 557. , Cham: Springer International Publishing, 2017, p. 900-910. doi: 10.1007/978-3-319-53480-0_89.
17. A. A. Freihat, M. Abbas, G. Bella, et F. Giunchiglia, « **Towards an Optimal Solution to Lemmatization in Arabic** », *Procedia Computer Science*, vol. 142, p. 132-140, 2018, doi: 10.1016/j.procs.2018.10.468.
18. S. Raharjo, R. Wardoyo, et A. E. Putra, « **Rule based sentence segmentation of Indonesian language** », *J. Eng. Appl.* Sci, vol. 13, no 21, p. 8986-8992, 2018.