Cross-linguistic Representation for Multilingual Access to Archaeological Data

Giulia Speranza, Innovative Industrial PhD student [gsperanza@unior.it]
Supervisor: Prof. PhD Johanna Monti [jmonti@unior.it]
UNIOR NLP Research Group, University of Naples “L’Orientale”
Department of Literary, Linguistic and Comparative Studies, Via Duomo 219 - 80138, Naples, Italy

1. Introduction

The project can be framed within the more general process of digital innovation of Cultural Heritage (CH), in particular, the internationalization of archaeological data in a multilingual perspective with the final aim of enhancing the interaction and communication between cultural institutions and the public, thus contributing to make the cultural experience in museums or archaeological sites more interactive and inclusive, overcoming linguistic and cultural barriers. The project is in line with the objectives set by the European Institutions for the Year of Cultural Heritage 2018, which promote the application of digital technologies in the field of humanities (Digital Humanities). The scientific community stresses the need for multilingualism in the field of CH, which, compared to other specialized field such as medicine or law, relies more and more on the use of the most common language to solve the issue of communicating globally [1]. The actual contribution that disciplines such as linguistics and machine translation (MT) may offer to cultural institutions, constitutes a valid method to provide international visitors with an inclusive cultural experience, bridging the linguistic gap.

2. State-of-the-art

The starting point of this research is to analyze the multilingual resources available online as open data such as dictionaries, thesauri, term banks (e.g. Getty and Wikidata). Another key-point is the investigation about previous research and applications in the field of computational linguistics and machine translation technologies applied to CLIR [2]. The research project also focuses on examining the recent experiments on Neural Machine Translation systems, with a special focus on domain-specific translation environments. Part of the project will concern an in-depth study about the possibilities offered by the semantic web technologies [3] and the application of LOD principles as well as CH standards (e.g. CIDOC, EDM) with particular attention to terminology and lexicography models for the representation and creation or conversion of annotated multilingual resources as LOD.

3. Methodology

The focus is on a multidisciplinary approach, which ranges from computational linguistics (CL) and Natural Language Processing (NLP) to the use of new technologies for machine translation (MT) in the field of Cross-Language Information Retrieval (CLIR), as well as the application of Linguistic Linked Open Data (LLOD) principles. An in-depth study of ontologies and the CIDOC Conceptual Reference Model (CRM) for describing the relationships used in cultural heritage documentation, is a crucial aspect of the research project.

4. Objectives

The final aim of this research is to create a digital repository containing information about cultural objects of archaeology in Italian, English and German, with the possibility of expanding it with other languages. Based on the gold terminology collected, a multilingual web portal or application for pc, smartphone and tablet will be developed. In conclusion, the resource to be implemented is intended to engage with a multicultural public making use of multimodal channels.

5. Case Study

The case study will be on the specialized terminology used to describe archaeological finds coming from the archaeological area of Campi Flegrei in the Campania region (Italy). The source data-set is composed of highly technical Italian terms but also words from everyday language, multi-word expressions (MWEs), Greek and Latin words, compounds, loan-words, cultural-bound elements and ambiguous words. This variety becomes even more compelling from a translation perspective. Multi-word expressions, in particular, are one of the trickiest issues when dealing with MT and NLP tasks [4].

6. Partners

- Smart Apps Soc. Coop.: provides its digital repository containing terminological dictionaries related to 27 types of CH in the Italian language (180,000 entries) and application programming skills
- Berlin School of Library and Information Science (Humboldt University of Berlin): offers insights for investigation on CLIR, Computer Science and Digital Cataloguing, European Data Model
- University of Naples “L’Orientale”: supplies ways for deepening knowledge on Linguistics, Computational Linguistics, Neural Machine Translation, Terminology management and Specialized Lexicon representation

7. References