A identificacão e referenciacão de entidades geográficas mencionadas: o caso da 'Peregrinação', de Fernão Mendes Pinto

Título A identificacão e referenciacão de entidades geográficas mencionadas: o caso da 'Peregrinação', de Fernão Mendes Pinto

Autor/a Afonso Xavier Canosa Rodríguez

Directores Pablo Gamallo Otero

Tipo Tese doutoral

Data de lectura 30/11/2017

Lugar de lectura Universidade de Santiago de Compostela

Abstract Geographical named entities represent one of the main types of named entities. A problem arises when a geographical named entity is identified in text but there are no given coordinates to provide a location. This thesis proposes a semantic model as a solution. Entities can be divided in two groups following an epistemologic criterion: those with known coordinates and those without. Peregrinação, an extensive report written by a diplomat travelling through Asia in the fifteenth century, is used as a case study. A list of geographical named entities is manually extracted and commented through comparative critical analysis of descriptions in corpus, those from related bibliography, and geovisualization of relevant areas in geographical databases and programs. This list is also used to evaluate automatic solutions for annotation and geo-referencing. Annotation is examined in three stages: tagging entities by matching expressions, optimization of results with a NERC tool and, finally, full automatization from scratch. For geo-referencing, entities with known coordinates are linked to an open global database from where geographical data is extracted and added to a local relational data-base. Relative references are solved for both known and unknown entities. The problem of assigning a geographical type is related to that of creating a taxonomy. For that purpose, extraction of geographical terms is evaluated, achieving best results by combining syntactic parsing, TF-IDF metrics and validation with external resources. A machine learning approach is explored to find examples of relations among entities and geographical features, results being significant for those entities with highest frequencies. Entities are organized in an onthology to refine their relations. An index is finally extracted to provide a structured definition of each entity, its occurrences in corpus, contemporary name and coordinates when available, and relations with other entities to further develop the relative reference.

LIGAZÓNS

Teseo

DESCARGAS

Referencia BibTex

Descargar versión completa

PROGRAMAS CIENTÍFICOS

Intelixencia de negocio e na web (antigo)