Solving unknown geographical named entities through semantic relations: the case of Fernão Mendes Pinto's *Peregrinação*

Afonso Xavier Canosa
University of Santiago de Compostela
canosarodrigues@gmail.com

1 Geo-referencing known and unknown geographical named entities

A typical approach for place-name geo-referencing takes an annotated geographical named entity, looks it up in a gazetteer and disambiguates candidates through textual and spatial analysis using either data inferred from the corpus or external resources, or both. The geo-reference is considered solved when its coordinates are found. However, a place-name may be unknown in terms of a precise location and yet spatial relations could be derived from its textual co-occurrences to narrow down the number of possible referents. This type of spatial information is particularly useful when the text has a high rate of unknown named entities that are difficult to retrieve in gazetteers, as it is the case of Fernão Mendes Pinto's *Peregrinação*, a collection of travels in Asia written in Portuguese in the 16th century.

2 Methods and results

With the aim to study the geographical value of Pinto's travels, entities with known coordinates were linked to an open global database (Geonames) to retrieve further geographical data. Those place-names that could not be located were processed as relative to a known entity. All entities were assigned to a geographical type and organized in an ontology to refine their relations. As one of the final products, a web environment processes the corpus and databases to provide a structured definition of each entity, its occurrences in the corpus, a contemporary name and coordinates when available, and relations with other entities (at least its parent, though also spatial relations of the type Distance_to when available). For annotation, standard statistical and rule-based NERC tools were applied, achieving significant results that justify automatic annotation as the starting point, though manual revision is needed to achieve the quality of a gold standard. Geo-referencing was done by expert research. Two semantic relations were considered in the ontology: hyponymy (to solve the geographical type) and meronymy (to solve the parent entity). Machine learning approaches were explored to find examples of relations among entities and geographical features, results being significant only for those entities with highest frequencies. A key question has been identified:

3 Question for further work

Working with historical texts written in non-standard language limits not only the availability of tools for NLP, but the use of machine learning methods is also challenged by the statistical relevance of the phenomena under study. In order to solve a semantic relation by means of quantitative methods as in distributional semantics, a certain threshold of occurrences has to be reached. Is a rule-based model the only (or the most efficient) solution?

14th International Conference on Spatial Information Theory.
Editors: NN; Article No. NN; pp. NN:1–NN:1
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany