Toponym Resolution using a learning algorithm in Ancient Upper Mesopotamia

Martin Unold
i3mainz
martin.unold@hs-mainz.de

Frank Boochs
i3mainz
frank.boochs@hs-mainz.de

Christophe Cruz
Le2i
christophe.cruz@u-bourgogne.de

Kai-Christian Bruhn
i3mainz
kai-christian.bruhn@hs-mainz.de

In the context of the project TexTelSem (http://higeomes.org), toponym resolution is the art of locating a place name (toponym) mentioned in an ancient upper Mesopotamian text. In this project, a geographical information system was developed and the toponyms with localization are available on the platform (http://checksem.u-bourgogne.fr/textelsem/TTS.html).

The semantisation of both databases ArchiBab and ArchiMass respectively from the Collège de France and the Free University of Berlin, and the archaeological data from the Ludwig Maximilian University Munich, is used to generate the maps of localized toponyms (e.g. old Babylonian, media-Assyrian and contemporary). The localization of ancient toponyms is possible due to the mapping between ancient toponyms and contemporary toponyms (e.g. archaeological find spots). However, some ancient toponyms do not have an identified clearly mapping with the available knowledge. Consequently, a learning approach is proposed to determine possible candidates for mapping.

A typical approach in computer science is to look for collocation frequency of typonyms in texts. More complex applications include also other terms within the textual context and properties of places. Toponyms come from surviving texts, and archaeological find spots are the potential Geo-locations for mapping. The amount of knowledge is limited, i.e. common algorithms in the area of toponym disambiguation (statistical methods, clustering, etc.) are not applicable.

The idea is the following. On one hand, every toponym gets a set of properties gained from the ancient texts, e.g. ”is a capital”, ”is located by a river” or ”has strong defense”, and cardinal directions to another toponym. On the other hand, archaeological findspots offer potential locations for Toponyms and Archaeological reports describe, what has been found in a certain place. Every findspot gets a set of properties, too. These properties are the findings described in the reports, e.g. ”walls”, ”ceramics” or ”cemetery”. In addition, there are already some resolved toponyms within the area under consideration to train algorithms. Unfortunately, the amount of resolvable toponyms is not enough to use common statistical methods and the data is full of uncertainty and incompleteness. The approach proposed consists to train of a learning algorithm. Since several properties rather occur in combination, the properties are not considered separately, but in combination. It means that the analysis works on the power set of properties and not on the properties themselves.

To avoid the influence of incompleteness and to make use of only few resolved toponyms, the statistic

includes also similar combinations of properties with a certain weight. This weight is determined by the similarity of property sets, e.g. by a Jaccard Index or by a Levenshtein Distance. The choice of the weight depends on the application. In addition, the algorithm searches the most promising fitting of an unresolved toponym to a certain find spot by calculating the probability of a fit depending on properties. This calculation is very costly, therefore the algorithm will only approximate the ”best” fitting by using techniques of non-linear optimization. One possibility to check the quality of this approach, is to remove one or several resolved toponyms from the training step and test, if the algorithm finds the correct resolution for those missing toponyms.

The application on the data from upper Mesopotamia with only a few properties shows the potential of this method. The method is tested by removing resolved toponyms and check, whether the algorithm detects the removed connection. The quality of the results is highly dependent of the quality and amount of information regarding toponyms and find spots. The approach could give some valuable hints in a lot of cases, but it is only to decision support and cannot replace human skills.