Preparing a RAMP on Location Name Resolution

Jean-Daniel Fekete, Aviz, Inria. Saclay, Dec. 2nd 2017

Many applications need to resolve lists of location names to actual locations. Typically, in Digital Humanities, visualizing on a map the locations of all the places mentioned in a document (book, letter, etc.) allow to quickly understand the geographical scope of the document. Additionally, resolving locations (latitude, longitude) mentioned in documents is a task performed routinely by many practitioners, such as archivists (trying to make sense and index of the contents of documents they manage), historians, and sociologists to name a few.

Recent systems allow mapping a precise location name to a location, thanks to Geocoging services (e.g. Google Map geocoging API[1]). However, these services are meant for modern location descriptions that are not ambiguous. Most documents contain location names that are ambiguous in the sense that multiple places have the same name, and even more have a similar name. We are interested in resolving **all** the location names contained in one document, and we assume that there is some level of consistency in these names that can be exploited by an automated tool. Systems such as CLAVIN[2] are made for resolving groups of location names using multiple heuristics. However, it is not clear how CLAVIN or other tools perform with different types of documents, such as books, letters, diaries, from different historical periods, or from different languages.

We therefore want to organize a RAMP (Rapid …) to study the performance of NLP/ML tools to address this problem: given a document where location names have been recognized, find the geographical location related to these names with the best accuracy.

To perform this RAMP, we will build a Web-based software to manually tag documents, and resolve names using existing systems. This software will allow us to create labels datasets that we will use for the RAMP. We will provide 20% of the resolved locations for each dataset we will provide (corpora) and challenge systems to resolve all the tagged location names.

We will devise a quality measure, based on the sum of distances to the right results and number of names actually resolved.

We will provide multiple corpora, some in English, some in French, some modern, some older (20th Century, 18th Century, 16th Century), and of various types (e.g. administrative, archival, letters, books). The RAMP should score programs by corpus, provided that some programs will only deal with English or French, or give-up on misspelled names.

We are collaborating with a research team in Toronto, Canada. They want to organize the RAMP in parallel in Toronto while we do it in Saclay. They have sponsors to provide prizes, if the CDS allow it (Thomson Reuters).

The resource we need is an intern/person for 1 month (or 2 months half time) to label the corpora using our prototype system. The person will also help us improve our interactive geocoding system that should, eventually, assist algorithms to resolve 100% of location names found in complex documents. The person should cost 3500€.

---

[1] https://developers.google.com/maps/documentation/geocoding/intro

[2] D'Ignazio, Catherine, et al. "Cliff-clavin: Determining geographic focus for news." NewsKDD: Data Science for News Publishing, at KDD 2014 (2014).