Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles

Dirk Ahlers[1]

Search Consultant, Tegucigalpa, Honduras,
ahlers@dhere.de

Abstract. Different language versions of Wikipedia contain articles referencing the same place. However, an article in one language does not necessarily mean it is available in another language as well and linked to. This paper examines geotagged articles describing places in Honduras in both the Spanish and the English language versions. It demonstrates that a method based on simple features can reliably identify article pairs describing the same semantic place concept and evaluates it against the existing interlinks as well as a manual assessment.

Keywords: Geospatial Web Search, Data fusion, Cross-lingual Information Retrieval, Record Linkage, Entity Resolution, Wikipedia, Honduras

1 Introduction

Wikipedia includes a vast amount of articles about places, many of which include a geographic coordinate that locates the content in the real world. Cross-language links are possible between articles describing the same place in multiple languages. Ideally, any article about a place of interest would include its correct coordinate and links to other Wikipedia language versions. This poses the question of how we can identify geotagged Wikipedia articles that describe the same place across different languages and what are useful similarity measures? The frame for this work is a country-specific search engine for the Latin American country of Honduras [1]. The geotagged wikipedia articles, together with gazetteer data (e.g., from geonames.org) can serve as initial knowledge about places and placenames.

The official language of Honduras is Spanish, and normally, the articles from this language would be expected to sufficiently cover the country. However, much information about the country comes from outside [4]. We therefore also looked into English articles to see if these would increase the coverage. This prompted the discovery of an interesting anomaly: Honduras has more English geotagged articles than Spanish ones. [1] Therefore, our aim is to merge both language versions and identify identical places on the article level. We analyze features and provide cross-language translations to define a similarity measure. This results

[1] Such anomalies exist in many countries: http://www.zerogeography.net/2012/10/dominant-wikipedia-language-by-country.html

Table 1. Examples of sibling articles

Spanish title	English title
Tegugigalpa	Tegucigalpa
Aeropuerto Internacional Toncontín	Toncontín International Airport
Santa Bárbara (Santa Bárbara)	Santa Bárbara, Honduras
El Paraíso	El Paraíso Department
Departamento de Copán	Copán Department
Virginia (Honduras)	Virginia, Lempira
Parque nacional Pico Bonito	Pico Bonito National Park

in an entity resolution algorithm for cross-lingual articles, and we share observations about the characteristics of Honduran locations in Wikipedia.

The approach we are following is called, varyingly, record linkage, entity fusion, entity resolution, or duplicate detection. [8] gives an overview on geospatial entity resolution. [7] use Wikipedia to ground and disambiguate place names. Merging geonames data to Wikipedia, [5] use a simple approach that if more than one entity exists in geonames with the same name, the closest entity within a distance of 5km is chosen. [6] add a translation approach to improve the title matching. This work is closest to ours, however, is still lacking in a graded consideration of both textual and positional similarity.

2 Wikipedia Language Fusion

We define the data fusion method in terms of finding language *siblings*. We combine text- and entity-based merging methods with geographic conflation techniques. For each article, we select and rank candidate siblings in the respective other language. The merging is based on the title and the location as shown in Table 1. The geographic feature type is rarely present, so it can only used as second-level evidence. For a comparison of two potential siblings, there are four cases to consider, 1. Names and coordinates match, 2. Names match, coordinates do not match, 3. Names do not match, coordinates match, 4. Names do not match, coordinates do not match. The first case is obviously trivial. All other cases are modeled by similarity measures based on non-exact matching.

Coordinates can vary due to different interpretations of the center of an area or variations in user-generated coordinates, especially for larger entities [3]. We limit the amount of candidate siblings we have to examine by cutting off the location similarity with a perimeter of 10km around an article's *location*, inside of which all candidates are examined.

For all candidates' *titles* within the radius, three cases would constitute a match, 1) titles match exactly, 2) titles match with small variations, 3) title can be translated and transposed to match. We define a title translation distance TTD as an editing distance similarity measure based on partial translations and permutations. The first case is easy, the second case only needs to account for spelling variations, which we do with a Levenshtein editing distance adapted

with a weight relative to the term length and with a reduced penalty for accents and tildes. Interestingly, most proper nouns are identical or very similar in both languages and can be well accounted for with the adapted Levenshtein distance. However, common nouns have to be translated and the order of terms within a placename also be changed. The translation table was filled mostly with relevant geographical feature types, taken from geonames (e.g., airports, islands, mountains, stadiums, cities, parks, etc.). Heuristics were generated about some conventions that we observed for both languages. For example, for municipality and department names, *Santa Bárbara (Honduras)*$_{ES}$ puts the higher-level administrative body, in this case the country name, in brackets, while *Santa Bárbara Department*$_{EN}$ uses the administrative type without a hint towards the country. This is helpful as often, departments and capital cities have the exact same coordinates.

To cover permutations, we employ a list of transposition heuristics as part of the translation. The inverted-first-pair translation swaps the first two terms: *Congreso Nacional de Honduras*$_{ES}$ $\rightarrow$ *National Congress of Honduras*$_{EN}$. The inverse order translation swaps first and last terms: *Río coco*$_{ES}$ $\rightarrow$ *Coco river*$_{EN}$; and the inverted-first-pair-moved translation inverts the order of the first two words and moves them to the end: *Parque nacional Pico Bonito*$_{ES}$ $\rightarrow$ *Pico Bonito National Park*$_{EN}$. We generate all potential variations of the title, including translations, and chose the variation with the minimum TTD and the smallest location distance as a sibling.

3 Evaluation

Honduras had 342 Spanish and 405 English articles, an 18% English overrepresentation. We use the wikipedia language interlinks as a ground truth for the evaluation. For all articles, the Wikipedia page and its interlinks were manually examined to determine siblings.

The algorithm resulted in 317 article pairs, 25 only Spanish articles, and 88 only in English (Fig. 1). Of these, 99.4% are correct pairs [2]. The articles without siblings are 84% correct, with 16% false negatives. Only two pairs were false positives. The first wrongly identifies *Comayaguëla*$_{ES}$ and *Comayagua*$_{EN}$ because they have both the exact same coordinates, even if the cities are about 80km apart. In this case the error lies with the incorrect coordinate in the article. The second assigns the department *Comayagua*$_{ES}$ to the city *Comayagua*$_{EN}$, which surprisingly is also wrong in the interlinks. This induces a subsequent error in the false negatives: *Comayagua (ciudad)*$_{ES}$ and *Comayagua Department*$_{EN}$ each had no siblings, but should have been matched to the previous pair. The other false negatives concern mostly slight mismatches paired with distanced coordinates, but also some more debatable ones, such as *Roatán (municipio)*$_{ES}$ and *Coxen Hole*$_{EN}$. When mapping articles as shown in Fig. 2, we see no language dominating certain regions but both languages distributed rather similarly.

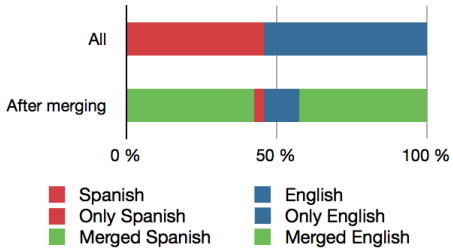

Fig. 1. Results of merging

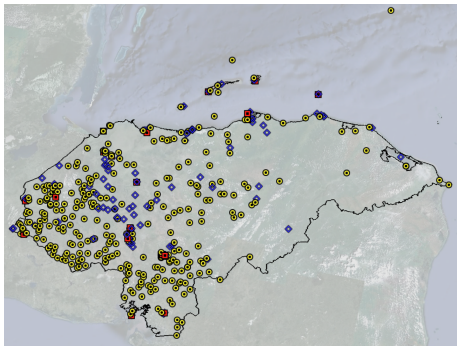

Fig. 2. Mapping of geotagged Wikipedia articles, yellow ⊙: merged articles, red ⊡: Spanish, blue ◊: English

4 Conclusion

The cross-language article resolution approach works on only simple features of location and title. It shows a good performance of 99.4% precision and 84% recall compared to a manually generated ground truth. The informally standardized titles of places such as municipalities versus capitals and the translation of geographic features drive the heuristics.

We expect the work to be transferable to other countries or language pairs. The country-dependent title-heuristics can be easily adapted. However, the approach of domain-specific translation of feature types and the 'implicit translation' by editing distance for the remaining terms mandates language pairs that are similar in both alphabet and spelling.

References

1. D. Ahlers. Towards Geospatial Search for Honduras. In *LACNEM 2011*, 2011.
2. D. Ahlers. On finding cross-lingual article pairs. *Tiny ToCS*, 1, 2012.
3. D. Ahlers and S. Boll. On the Accuracy of Online Geocoders. In *Geoinformatik 2009*, 2009.
4. D. Ahlers, J. Matute, I. Martinez, and C. Kumar. Mapping the Web resources of a developing country. In *GI Zeitgeist 2012*, 2012.
5. J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal*, 2012.
6. Y. Liu and M. Yoshioka. Construction of large geographical database by merging Wikipedia's Geo-entities and GeoNames. Tech Report SIG-SWO-A1102-03, 2011.
7. S. E. Overell and S. M. Rüger. Identifying and grounding descriptions of places. In *GIR 2006*, 2006.
8. V. Sehgal, L. Getoor, and P. D. Viechnicki. Entity Resolution in Geospatial Data Integration. In *GIS'06*, 2006.