Toponym Disambiguation using Ontology-based Semantic Similarity

David S Batista[1], João D Ferreira[2],
Francisco M Couto[2], and Mário J. Silva[1]

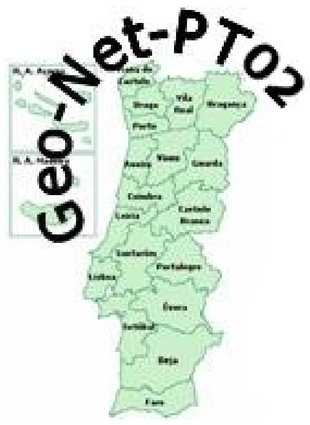
[1]IST, INESC-ID, Technical University of Lisbon
[2]LaSIGE, University of Lisbon

Geographic ambiguity

Referent ambiguity, the same toponym can represent more than one geographic concept.

"I was born close to *Santa Catarina* in *Lisboa*."

which *Lisboa and Santa Catarina is the* sentence referring to? According to Geo-Net-PT02, *Santa Catarina* can represent up to 104 different locations and *Lisboa* 41, from streets to a municipality, a city or a region. Geo-Net-PT02 is a public geographic ontology covering the territory of Portugal. It is divided in two domains: administrative and physical. The ontology is structured as a directed acyclic graph (DAG).



Semantic Similarity

If an ontology is structured as a DAG semantic similarity measures based on the information content (IC) that two concepts share can be applied.

$$\mathrm{IC(c)}=\text{-}\log \frac{f(c)}{\max_c f(c)}$$

$$\mathrm{IC}_{\mathrm{MICA}}(c_1, c_2) = \max\{\mathrm{IC}(a) : a \in \mathrm{Anc}(c_1) \cap \mathrm{Anc}(c_2)\}$$

Jiang and Conrath defined a distance measure as the difference between the IC of both concepts and the IC of their MICA. Assuming that the IC is normalized for values between 0 and 1, the distance can be converted to similarity:

$$\mathrm{sim}_{\mathrm{JC}}(c_1, c_2) = 1 - (\mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 \times \mathrm{IC}_{\mathrm{MICA}}(c_1, c_2))$$

Lin defined similarity as the IC of their MICA over the IC of both concepts:

$$\mathrm{sim}_{\mathrm{Lin}}(c_1, c_2) = 2 \times \mathrm{IC}_{\mathrm{MICA}}(c_1, c_2) \div (\mathrm{IC}(c_1) + \mathrm{IC}(c_2))$$

Resnik defined similarity between two concepts as the amount of information content they share, given by the information content of their MICA:

$$\mathrm{sim}_{\mathrm{Resnik}}(c_1, c_2) = \mathrm{IC}_{\mathrm{MICA}}(c_1, c_2)$$

Each measure gives a score [1,0] reflecting the geographic closeness according to the ontology between the two concepts.

Toponym Disambiguation

Having as input a sequence of toponyms, extracted, for instance, from a text:

$$T = \{t_1, \ldots, t_n\}$$

we define for each toponym, the set of geographic concepts labeled with the toponym as:

$$GeoConcepts(t_x) = \{g_1, \ldots, g_n\}$$

the goal is to define a function that maps each toponym to the geographic concept it is intended to represent in the input sequence:

$$GeoMap(t_x) = g_x : g_x \in GeoConcepts(t_x)$$

Global-Mapping identifies for each toponym the concept that maximizes its semantic similarity with the concepts for all the other toponyms.

$$GeoMap_{\mathrm{global}}(t_x) = \arg\max_{g_x} (\max_{g_y} \ sim(g_x, g_y))$$

where: $g_x \in GeoConcepts(t_x)$ and $g_y \in GeoConcepts(T \backslash \{t_x\})$

Each toponym t_x is mapped to the unique geographic concept that has the highest similarity score among all pairs of geographic concepts.

Sequential-Mapping takes into consideration the order of the toponyms in the text. First, it calculates the semantic similarity between the pairs of concepts for the first pair of toponyms, t_1 and t_2 :

$$GeoMap_{\mathrm{seq}}(t_1, t_2) = \arg\max_{g_1, g_2} \ sim(g_1, g_2)$$

Then, the next toponym in the text, t_3 is disambiguated by the geographic concept that gave the highest similarity score to toponym t_2 and all the possible geographic concepts for the toponym t_3 the pair with the highest semantic similarity is chosen.
This pattern is applied sequentially, until the last toponym is reached. This technique ensures that the geographic concept that yields the maximum similarity is always propagated to the next pair:

$$GeoMap_{\mathrm{seq}}(t_x : 3 \leq x \leq n) = \arg\max_{g_x} \ sim(GeoMap_{\mathrm{seq}}(t_{x-1}), g_x)$$

Assessment

The Information Content (IC) of the geographic concepts in Geo-Net-PT02 was calculated with basis on the number of occurrences of the capitalized version of the name of a concept in a Portuguese n-grams collection, WPT05.

As baseline for assessing the effect of the proposed mapping techniques and semantic similarity measures we applied a naive disambiguation technique that simply selects the geographic concept with the highest IC:

$$GeoMap_{\mathrm{baseline}}(t_x) = \arg\max_{g_x} \ IC(g_x)$$

We then applied the three techniques to automatically map the toponyms from Geo-CHAVE-PT to geographic concepts in Geo-Net-PT02. To evaluate the mappings we applied a formula, *GeoSimilarity(g1,g2)* to measure the geographic similarity between two concepts, where *g1* represents the geographic concept manually mapped and *g2* the concept automatically disambiguated.

Geo-CHAVE-PT

Geo-CHAVE-PT is a Portuguese collection of news articles, with toponyms which are part of the Portuguese territory and having a geographic concept in Geo-Net-PT annotated. The collection as a total of 195 news articles of different categories published between 1994 and 1995.

Categories	Articles	#Geographic Entities
Local	124	972
National	35	218
Society	14	124
Diverse	3	26
Economy	4	21
Sport	4	17
Science	4	24
Culture	7	61
Total	195	1463

Table 1: news articles categories in Geo-CHAVE-PT

Geographic Entity Type	Percentage
Localidade	52.90%
Concelho	17.29%
Freguesia	7.38%
País	6.70%
Zona	1.98%
Rua	1.98%
Distrito	1.71%
Praça	1.64%
NUT2	1.64%
Província	1.44%
Avenida	<1%

Table 2: diversity of geographic entity types in Geo-CHAVE-PT

Geographic Similarity

$$\mathrm{closeness}(g_1, g_2) = \frac{1}{1 + \mathrm{shortestpath}(g_1, g_2)}$$

$$\mathrm{relatedness}(g_1, g_2) = \begin{cases} \mathrm{desc}(g_1) \div \mathrm{desc}(g_2) & \text{if } g_1 \subseteq g_2 \\ \mathrm{desc}(g_2) \div \mathrm{desc}(g_1) & \text{if } g_2 \subseteq g_1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{sibilings}(g_1, g_2) = \begin{cases} 1 & \text{if } \mathrm{parent}(g_1) = \mathrm{parent}(g_2) \\ 0 & \text{otherwise} \end{cases}$$

$\mathrm{desc}(g_x)$ is the number of descendants of g_x in the graph

$\mathrm{shortestpath}(g_x, g_y)$ defines the minimum distance between g_x and g_y measured in number of edges

$\mathrm{parent}(g_x)$ is the concept in the ontology hierarchy immediately above g_x

Those concepts are then combined by a sum and normalized to [0, 1], yielding the metric adopted for this study:

$$GeoSimilarity(g_1, g_2) = \frac{(\mathrm{closeness}(g_1, g_2) + \mathrm{relatedness}(g_1, g_2) + \mathrm{sibilings}(g_1, g_2))}{3}$$

Results

Technique	Similarity Measure	Average $GeoSimilarity$	CPU time
$GeoMap_{\mathrm{seq}}$	Jiang-Conrath	0.54	01:02:20
	Lin	0.45	01:03:45
	Resnik	0.43	03:04:57
$GeoMap_{\mathrm{global}}$	Jiang-Conrath	0.51	02:17:27
	Lin	0.43	02:18:45
	Resnik	0.43	02:18:47
$GeoMap_{\mathrm{baseline}}$		0.28	00:01:12

Conclusions

The Jiang-Conrath semantic similarity measure yields the best results and both mapping techniques have comparable results. The Global-Mapping technique, however, has high computational costs and assumes one-sense-per-word, the Sequential-Mapping is faster, and allows repeated toponyms in the same text to be correctly mapped to different geographic concepts. The extraction of toponyms did not take into consideration linguistic features such as sentence boundaries or paragraphs. Geographic features usually associated to a toponym, such as municipality (*concelho*), street (*rua*), were not taken into consideration. Such geographic features alone can disambiguate the toponyms. Combining this new heuristic with others can also improve the geographic mapping process.

All the resources used in this work are publicly available:

Geo-CHAVE-PT http://dmir.inesc-id.pt/reaction/Geo-Net-PT_02_in_English
Geo-Net-PT02 http://linguateca.pt/geonetpt/geonetpt02/
WPT05 http://dmir.inesc-id.pt/reaction/WPT_05_in_English

This work was supported by Fundação para a Ciência e Tecnologia (FCT), through the financial support of the REACTION project (UTA-Est/MAI/0006/2009) and the PhD scholarships SFRH/BD/70478/2010 and SFRH/BD/69345/2010, and the PIDDAC Program funds (INESC-ID multi annual funding) and through the LASIGE multi annual support.