

# What Kinds of Geographical Information Are There in the Portuguese Web?

Marcirio Silveira Chaves<sup>1</sup> and Diana Santos<sup>2</sup>

<sup>1</sup> Node of XLDB at University of Lisbon, Linguatca

<sup>2</sup> Node of Oslo at SINTEF ICT, Linguatca

**Abstract.** This paper presents some results about the geographical information in the Portuguese web and the overlap with people's and organization's named entities, using a geographic ontology based on authoritative sources and a named entity recognizer.

## 1 Introduction

This paper describes ongoing work on the study of a snapshot of the Portuguese web to assess how geographical information is encoded in the web in Portuguese (note that we are dealing here with a different web community and in a different language). We started by identifying and classifying named entities (NEs) focusing on locations.

Although place names are the kind of named entities which are easier to identify with the help of gazetteers (at least for English newspaper text [1]), the lack of a good named entity recognition (NER) system may hamper attempts to (naively) assign geographic scopes to web documents. For example, Zong et al. found that the ambiguity between geographic and non-geographic names was the main cause of errors [2].

## 2 Context and Available Resources

In order to conduct our study, we used several freely available resources, which we briefly describe in turn: in the GREASE project [3], associated to the tumba! search engine (`tumba.pt`), the first author and colleagues have created the Geographic Knowledge Base (GKB) [4]. GKB integrates items from multiple administrative information sources plus Internet data such as names of sites and domains in Portugal. Information in GKB about Portugal is exported as an ontology named Geo-Net-PT01 (`xldb.fc.ul.pt/geonetpt/`). Geo-Net-PT01 contains ca. 400,000 geographic terms and ca. 400,000 geographic relationships.

In addition, we used a Portuguese web collection called WPT 03 (`linguateca.di.fc.ul.pt/WPT03/`), with 12 Gbytes and 3.7 million pages comprising 1.6 billion words. Roughly 68.6% of these pages are in Portuguese.

In our experiments, we used the SIEMÊS NER system [5]. In the evaluation contest of NER for Portuguese, HAREM [6], SIEMÊS scores for the location category were close to 70% of precision and 75% recall. However, the version of SIEMÊS used in the our experiments is different from that used in HAREM.

### 3 Getting to Know Better the Geographic References

The result of tagging the first randomly selected 32,000 documents with SIEMÊS is summed up in Table 1. We chose the three following categories of NE's: People, Organizations and Locations. SIEMÊS was configured to assign the People and Organizations because they are frequently ambiguous with or related to location terms, respectively. In Portugal, there are several surnames identical to location names, as in "Irene *Lisboa*" or "Camilo *Castelo Branco*." Also, we wanted to investigate how often a location was included in the name of an organization, to estimate how many cases it does provide a reliable clue to the physical place the organization is located in.<sup>1</sup>

**Table 1.** NEs detected in a 32,000 documents sample of WPT 03. MW stands for multi-word and GN stands for Geo-Net-PT01. DNE: Distinct named entities (types).

	# of NEs (%)	# of DNEs	# of MW NEs (%)	# of MW DNEs (%)	# of MW DNEs containing a name in GN (%)	# of DNEs occurring in GN (%)
PEO	250,585 (26.48)	77,228	140,155 (55.93)	58,991 (76.39)	24,105 (31.21)	521 (0.67)
ORG	418,915 (44.27)	114,353	214,698 (51.25)	89,790 (78.52)	26,789 (23.43)	462 (0.40)
LOC	276,775 (29.25)	47,972	90,018 (32.52)	36,395 (75.87)	22,959 (47.86)	4,576 (9.53)
Sum	946,275 (100.00)	239,553	444,871 (47.01)	185,176(77.30)	73,853 (30.83)	5,559 (2.32)

Table 1 shows that close to 1 million of NEs, belonging to the three categories, were identified, 30% of which corresponding to locations. For all categories, more than 75% of DNEs are multi-word. Organization names were the most frequent as far as tokens and types are concerned. As to type/token ratio, people NEs were the most varied, while locations displayed the lowest variation (i.e., location names were considerably more repeated in the sample than person names).

The last two columns of Table 1 measure partial and total overlap: while ambiguity with people's or organization's names is less than 1%, as much as 31.21% of the person DNEs and 23.43% of the organization DNEs contain a geographic name included in Geo-Net-PT-01.<sup>2</sup>

As to overlap of geographic locations in the web and in Geo-Net-PT-01, the numbers are astonishing at first: considering that Geo-Net-PT-01 is supposed to be complete as to Portuguese administrative geography, why only around 10% of the distinct locations present in the Portuguese web should appear in Geo-Net-PT-01? Even taking into account spelling errors or non-official naming conventions it is hard to account for the other 90%.

<sup>1</sup> For example, *Universidade do Porto* entails that it is located in *Porto*, while *Associação de Amizade Portugal-Itália* has no relation with location (the name refers to friendship among the peoples of the two countries) and this is even less so in the case of *Pastelaria Finlândia*, a name for a pastry shop.

<sup>2</sup> For this comparison, we used all names in Geo-Net-PT-01 (27,855), except for street names and postal codes.

**Table 2.** Distribution of the types contained in the local (LOC) category

Type	# of DNEs(%)	# of MW DNEs(%)
POV (names of pop. places)	33,827 (70.51)	24,037 (71.06)
ENDRALAR (full address)	3,505 (7.31)	3,313 (94.52)
SOCCUL (society/culture)	3,474 (7.24)	3,161 (90.99)
PAIS (country)	1,987 (4.14)	1,419 (71.41)
RLG (religion)	1,197 (2.50)	1,113 (92.98)
Other ( $\sum$ 11 types)	3,982 (8.30)	3,352 (84.18)
Sum	47,972 (100,00)	36,395 (75,87)

**Table 3.** Distribution of NEs per document

	Total Distinct			Total Distinct	
Avg. PEOs. per doc. with PEOs.	11.65	7.82	Median LOCs	4	3
Avg. ORGs. per doc. with ORGs.	13.81	9.78	Stdev LOCs	149.7	57.54
Avg. LOCs. per doc. with LOCs.	11.31	7.34	# docs. with 1 LOC	5,443	6,184
Avg. NEs per doc. with NEs	30.04	20.47	# docs. > 3 LOCs	12,913	11,640
Maximum # of LOCs in 1 doc.	20,594	6,472	# docs. > 30 LOCs	1,483	713

To investigate whether the kind of location occurring in Portuguese web texts had different properties: more fine-grained, or relating to physical geography (rivers, mountains, etc.), we looked into the subtypes of location NEs provided by SIEMÉS, shown in Table 2. The most frequent type of geographic NE is the name of a city, town or village (POV), followed by the name of a country. In fact, more than 85% of LOCs are concentrated in just three types (POV, ENDRALAR and SOCCUL) and the same occurs when counting only multi-word names.

Our explanation for this wealth of location NEs not present in Geo-Net-PT-01 (in addition to a systematic overgeneration of SIEMÉS, which will have to be analysed elsewhere, although we expect it to be of considerable import for the numbers presented here) resorts to the following hypotheses:

- given that Portugal is and has always been a small country with a very worldwide perspective, many (or even most) of the web pages do not concern only (or specially) Portugal – and so many geographical named entities concern places in foreign countries;
- in texts, people are bound to write about more fine-grained locations (often deictically), as “downtown”, “near my old school”, “in front of Jerónimos”, or “in the A1 highway”, which are not part of an administrative ontology.

Finally, with this study we also wanted to assess the hypothesis that location is a transversal semantic category, in the sense that geographical information can be found in (almost) all sorts of texts and not only in specialized technical texts (as for example those dealing with geography or tourism). So, we measured the total number of documents with at least one NE: 31,489 (98.4% of the snapshot). References to people are present in 21,499 (67.18%) documents, organizations in

30,328 (94.77%) documents and locations in 24,468 (76.46%) documents. Table 3 shows that each document (containing at least one NE) contains on average ca. 20 DNEs from which more than seven are localities and ca. 50% of the documents with LOCs contain more than three LOCs. The values of the “Distinct” column measure the distinct NEs into each document.

## 4 Concluding Remarks

We provide a first measure of geographical information (as far as named entities are concerned) in the Portuguese web, concluding that a geographic ontology built from web texts can complement administrative sources.

Conversely, and using the available frequency lists for the WPT 03 collection, we found out that ca. 20% of the one-word names in Geo-Net-PT01 do not occur in WPT 03 at all. This shows that – for indexing of Web contents – to know what the Web talks about may considerably reduce the index size, and provides another motivation for our work.

## Acknowledgements

We are grateful to Mário Silva for pertinent comments and to Luís Sarmiento for help with SIEMÊS. This work was supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia, co-financed by POSI.

## References

1. A. Mikheev, M. Moens, and C. Grover. Named Entity Recognition without Gazetteers. In *Proc. of the Ninth International Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Bergen, Norway, 1999.
2. Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. On assigning place names to geography related web pages. In *JCDL '05: Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*,.
3. M. J. Silva, B. Martins, M. S. Chaves, N. Cardoso, and A. P. Afonso. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems, Elsevier Science*, 2006 (in press).
4. M. S. Chaves, M. J. Silva, and B. Martins. A Geographic Knowledge Base for Semantic Web Applications. In C. A. Heuser, editor, *Proc. of the 20<sup>th</sup> Brazilian Symposium on Databases, Uberlândia, Minas Gerais, Brazil*, pages 40–54, October, 3–7 2005.
5. Luis Sarmiento. SIEMÊS - a named entity recognizer for Portuguese relying on similarity rules. This volume.
6. Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: an Advanced NER Evaluation Contest for Portuguese. In *Proceedings of LREC'2006*, Genoa, Italy, 22-28 May 2006.