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 Lisbon, Portugal
{dsbatista,msilva}@inesc-id.pt
[2] University of Lisbon, LaSIGE

Abstract. We propose a new heuristic for toponym sense disambiguation, to be used when mapping toponyms in text to ontology concepts, using techniques based on semantic similarity measures. We evaluated the proposed approach using a collection of Portuguese news articles from which the geographic entity names were extracted and then manually mapped to concepts in a geospatial ontology covering the territory of Portugal. The results suggest that using semantic similarity to disambiguate toponyms in text produces good results, in comparison with a baseline method.

1 Introduction

Word-sense-disambiguation deals with the problem of selecting the correct semantic meaning of ambiguous words mentioned in an unstructured text. A particular case of ambiguous words are toponyms (place names), whose geographic meaning, that is the geographic concept identified by a unique place on the Earth to which the word is referring to, needs to be identified. For instance, the place name *Lisboa* can represent up to 41 different locations in the territory of Portugal alone, from streets to a municipality, a city or a region.

In this work we address one specific type of ambiguity, namely *referent ambiguity*, when the same toponym can represent more than one geographic concept. Given a toponym in a text and a geospatial ontology, we assign, from the set of concepts having that toponym, the one representing the toponym's context.

A common approach to resolve *referent ambiguity* uses heuristics, usually based on hierarchy constraints derived from administrative subdivisions encoded in the ontology [10]. For example, we expect that a large city is more likely to be referred than a small village with the same name. Another method uses discourse interpretation heuristics. For instance, if the same toponym is used multiple times in the same text or section, it is assumed that it is always referring to the same location rather than different locations that share the same name [5]. Yet another method, based on using geospatial information associated with each concept, is to minimize the bounding polygon that contains all candidate referents, using the geographic bounding boxes associated with each concept [7].

In this paper we introduce a new heuristic. It is reasonable to expect that in a section of a text, close geospatial concepts share a higher degree of closeness

also in semantics. For instance, if a news article mentions *Lisboa* and *Porto*, than its expected that *Lisboa* refers to the city and not to one of the streets with that name, since both these names can refer to cities. Based on this idea, we developed two mapping techniques based on semantic similarity measures to solve *referent ambiguity*: Global-Mapping and Sequential-Mapping. In the evaluation we used a geospatial ontology of the Portuguese territory and a set of geographic annotated news articles. Experiences show that Sequential-Mapping based on the Jiang-Conrath measure gives the best results. The rest of this paper is organized as follows: in section 2 we present the semantic similarity measures used, section 3 describes the two mapping techniques, section 4 the assessment and results, and finally in section 5 we present the conclusions.

2 Semantic Similarity

According to Budanitsky and Hirst, the most effective semantic similarity measures are the ones based on the information content (IC) that two concepts share [3].

If the ontology is structured as a directed acyclic graph (DAG) the IC of an ontology concept is inversely proportional to its frequency in a given corpus. The frequency is propagated to its ancestors, making the IC of a concept roughly proportional to its depth in the DAG. Thus, if $f(c)$ is the frequency of concept c, including its descendants, IC is defined as $\mathrm{IC}(c) = -\log \frac{f(c)}{\max_c f(c)}$, where $\max_c f(c)$ is the maximum frequency of all concepts, i.e. the frequency of the root concept. The shared information between two concepts is normally proportional to the IC of the Most Informative Common Ancestor (MICA) in the ontology:

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

where $\mathrm{Anc}(c_x)$ represents the ancestors of c_x.

Resnik defined similarity between two concepts as the amount of information content they share, given by the information content of their MICA [11]. Jiang and Conrath defined a distance measure as the difference between the IC of both concepts and the IC of their MICA [6]; assuming that the IC is normalized for values between 0 and 1, the distance can be converted to similarity. Lin defined similarity as the IC of their MICA over the IC of both concepts [8]. Every one of these measures is based on Resnik's definition of shared information (they use a single common ancestor), and are summarized in Table 1.

3 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:

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

Table 1. Semantic Similarity Measures

Measure	**Formula**
Jiang-Conrath	$sim_{\mathrm{JC}}(c_1, c_2) = 1 - (\mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 \times \mathrm{IC}_{\mathrm{MICA}}(c_1, c_2))$
Lin	$sim_{\mathrm{Lin}}(c_1, c_2) = 2 \times \mathrm{IC}_{\mathrm{MICA}}(c_1, c_2) \div (\mathrm{IC}(c_1) + \mathrm{IC}(c_2))$
Resnik	$sim_{\mathrm{Resnik}}(c_1, c_2) = \mathrm{IC}_{\mathrm{MICA}}(c_1, c_2)$

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. One-sense-per-word is assumed, that is, if the same toponym occurs in the text more than once we always assume that it is referring to the same geographic location [5]. For every toponym t_x a geographic concept is assigned:

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

where $g_x \in GeoConcepts(t_x)$ and $g_y \in GeoConcepts(T \backslash \{t_x\})$. At the end, each toponym t_x is mapped to the unique geographic concept that has the highest similarity score among all pairs of geographic concepts. This technique explores all the possible combinations between the different geographic meanings that each toponym can have, which represents a high computational complexity.

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. From the set of possible pairs, the one with the highest semantic similarity is chosen, and the two geographic concepts are mapped to the corresponding toponyms:

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

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

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

4 Assessment

To evaluate our techniques we used three public resources, described below.

Geo-Net-PT is a public geographic ontology covering the territory of Portugal. It is divided in two domains: administrative and physical. The administrative domain contains the administrative divisions of the territory and the physical domain includes physical geography features, such as natural regions and man-made spots [9].

The Information Content (IC) for any given concept 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 [2].

CHAVE [12] is a Portuguese collection of news articles, with toponyms recognized by REMBRANDT [4]. The articles were scanned to map each identified toponym to the geographic concepts it might represent in Geo-Net-PT. A total of 195 news articles were selected for manual mapping. From the set of possible geographic concepts associated to each toponym, human mappers discarded all but the one representing the correct geographic concept associated to the toponym's context. Only toponyms for parts of the Portuguese territory having a geographic concept in Geo-Net-PT were considered. The result is Geo-CHAVE-PT, a subset of news articles from CHAVE with the toponyms linked to Geo-Net-PT concepts. Geo-Chave-PT is available for download[3], along with a detailed description of the corpus and mapping guidelines.

As baseline for assessing the effect of the proposed semantic similarity measures, we applied a naïve disambiguation technique that simply selects the geographic concept with the highest IC: $GeoMap_{\text{baseline}}(t_x) = \arg\max_{g_x} \ IC(g_x)$

4.1 Assessing Geographic Similarity

We applied the three techniques to automatically map the toponyms to geographic concepts. To evaluate the mappings we adapted a previously proposed formula to measure the geographic similarity between two concepts [1]. For a given pair (g_1, g_2), where g_1 represents the geographic concept manually mapped and g_2 the concept automatically disambiguated, the following characteristics were calculated:

$$\text{closeness}(g_1, g_2) = (1 + \text{shortestpath}(g_1, g_2))^{-1}$$

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

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

where $\text{desc}(g_x)$ the number of descendants of g_x in the graph, $\text{shortestpath}(g_x, g_y)$ defines the minimum distance between g_x and g_y, measured in number of edges

[3] `http://dmir.inesc-id.pt/reaction/Geo-Net-PT_02_in_English`

Table 2. Average *GeoSimilarity* and processing time for the Geo-CHAVE-PT articles

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

and 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:

$$GeoSimilarity(g_1, g_2) = \frac{1}{3}(\text{closeness}(g_1, g_2)+\text{relatedness}(g_1, g_2)+\text{sibilings}(g_1, g_2))$$

4.2 Results

The semantic similarity measures were implemented in Java, querying a relational database representation of Geo-Net-PT. Each mapping technique, implemented in Python, processed the articles from Geo-CHAVE-PT as a batch job. The processing time and average *GeoSimilarity* for the whole collection are shown in Table 2.

The Jiang-Conrath measure applied to the Sequential-Mapping achieves the best results. It also takes less time to process, because it does not explore all the possibilities. It simply chooses the best one locally as it sequentially maps the toponyms to geographic concepts. The Resnik measure took three times more to process in the Sequential-Mapping, because the highest similarity score was too often the same for different pairs, probably because it only uses the IC of a single common ancestor, the most informative one, the MICA. This increased the number of disambiguation possibilities exponentially.

This result, for a geospatial ontology, is in line with the work of Budanitsky and Hirst, where the Jiang-Conrath measure also outperformed the other tested measures on WordNet [3].

5 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.

6 Acknowledgements

This work was supported by FCT, through the project UTA-Est/MAI/0006/2009 (REACTION), the scholarship SFRH/BD/70478/2010 and the Multiannual Funding Program.

References

1. L. Andrade and M. J. Silva. Relevance Ranking for Geographic IR. In *GIR-2006, the 3rd Workshop on Geographical Information Retrieval*, August 2006.
2. D. Batista and M. J. Silva. A Statistical Study of the WPT05 Crawl of the Portuguese Web. In *FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop*, Vigo, Spain, 2010.
3. A. Budanitsky and G. Hirst. Semantic distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Proc. of the Workshop on WordNet and Other Lexical Resources*, 2001.
4. N. Cardoso. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In *Encontro do Segundo HAREM, PROPOR 2008*, 2008.
5. W. A. Gale, K. W. Church, and D. Yarowsky. One Sense per Discourse. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, 1992.
6. J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th International Conference on Research on Computational Linguistics*, 1997.
7. J. L. Leidner, G. Sinclair, and B. Webber. Grounding Spatial Named Entities for Information Extraction and Question Answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, 2003.
8. D. Lin. An Information-Theoretic Definition of Similarity. In *Proc. of the 15th International Conference on Machine Learning*, 1998.
9. F. J. Lopez-Pellicer, M. Chaves, C. Rodrigues, and M. J. Silva. Geographic Ontologies Production in Grease-II. Technical Report TR 09-18, University of Lisbon, Faculty of Sciences, LASIGE, November 2009.
10. E. Rauch, M. Bukatin, and K. Baker. A Confidence-Based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*. Association for Computational Linguistics, 2003.
11. P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 1995.
12. D. Santos and P. Rocha. The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In *Multilingual Information Access for Text, Speech and Images*. 2005.