Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

Discovering the spatial coverage of the documents through the SpatialCIM Methodology

Rosa Nathalie Portugal Vargas, Solange Oliveira Rezende
University of São Paulo, Computer Science Department
Av. Trabalhador São Carlense, 400.
São Carlos-São Paulo, Brazil
{nathalie, solange}@icmc.usp.br

Maria Fernanda Moura, Eduardo Antonio Speranza, Ercilia Rodriguez
Embrapa Agricultural Information
Av. André Tosello, 209 – Barão Geraldo
Campinas-São Paulo, Brazil
{fernanda,speranza,erciliasr}@cnptia.embrapa.br

Abstract

The main focus of this paper is to present the SpatialCIM methodology to identify the spatial coverage of the documents in the Brazilian geographic area. This methodology uses a linguistic tool to assist in the entity recognition process. The linguistic tool classifies the recognized entities as person, organization, time and localization, among others. The localization entities are checked using a geographic information system (GIS) in order to extract the Brazilian entity geographic paths. If there are multiple geographic paths for a single entity, the disambiguation process is carried out. This process attempts to locate the best geographic path for an entity considering all the geographic entities in the text. Another important objective of this paper is to show that the disambiguation process improves the geographic classification of the documents considering the obtained geographic paths. The validation process considers a set of news previously labeled by an expert and compared with the results of the disambiguated and non-disambiguated geographic paths. The results showed that the disambiguation process improves the classification compared with the classification without disambiguation.
Keywords: Ambiguity problem resolution, spatial coverage identification, toponym resolution.

1 Introduction

In the last decades the internet has captured the attention of many users, not only for the services offered but also for the possibility to access a huge amount of information represented in different kind of formats like the on-line news. This huge amount of information is not easily interpreted, demanding tools and techniques which allow structuring and organizing the data. One common task is to browse texts by place name when searching for information about a specific event or location [1]; [2]; [3]. Therefore there is a need to automatically classify the documents with their respective geographic paths; which can be obtained with the support of text mining techniques. The text mining techniques aim to discover and extract innovative knowledge in textual collections [4].

There are many researches focused on the geographic localization of documents considering the implicit ambiguity problem. For example, the geographic entity "Paris'" can be mapped in "France" or in "United States". These researches usually work with the extracted geographic features of the document structures [5], images [6], hyperlinks [7] or ontologies [8], with the aim to determine more precisely the geographic zone where they belong to.

The toponym disambiguation process is the task to found the spatial localization in text for the entity (toponym) using a structured representation as geographic coordinates, information of data bases, or localization obtained from geographic ontologies [9].

In order to solve the ambiguity problem, techniques which allow the entity identification according to their characteristics were developed. The ambiguity can be understood as a word or phrase which has more than one meaning in the language to which the word belongs. For example, the phrase "Touristic places in Barcelona" represents an ambiguous entity since "Barcelona" can be located in Spain, Venezuela or Brazil. The most common techniques of disambiguation use the gazetteers in order to provide the geographic the coordinates of each entity. A gazetteer is a catalogue of locations or places (dictionary of toponyms) which provides a vocabulary of geographic terms along with their respective locations [10] [9].

This paper presents the Spatial Coverage Identification Methodology (SpatialCIM) that addresses the ambiguity problem, and uses a linguistic tool to assist in the entity recognition process. The topic of the set of news used in the experimental phase is the sugar cane in the Brazilian area. As a result of the news nature, the associated geographic paths are mapped into the Brazilian area considering the sugar mills location and the recognized international entities. The path representation follows the hierarchical structure of "Region, State, Meso-Region, Micro-Region, City, Sugar Mill and Category", according to the Brazilian territorial division.

The paper is organized as follows: Section 2 covers the related work; Section 3 introduces the proposed methodology to determine the spatial coverage of the news; Section 4 presents the preliminary experiments and the obtained results. Then, the final considerations of the work are presented in Section 5.

2 Related Work

In order to determine the geographic spatial coverage of the documents it is necessary to recognize the entities (toponym) in the text, since they are considered as an important source of geographic information. The identification of the toponyms is usually carried out with the help of gazetteers that provide additional geographic information like geographic coordinates, class (city, river, and country), population, and size, among others. In the process of recognizing the toponyms, different types of ambiguities can be found and they need to be solved in order to establish the correct geographic classification of the documents. According to [11] there are two types of ambiguities: (i) geo/non-geo, and (ii) geo/geo. A (i)geo/non-geo ambiguity occurs when an entity has a non-geographic meaning, such as the word "Turkey". A (ii) geo/geo ambiguity occurs when distinct entities have the same name as London in England and London in Ontario. There are many researches that try to solve the ambiguity problem in different ways.

In [12], Leidner proposed a set of heuristics that allows to solve the ambiguity problem. A heuristic known as "second minimalistic heuristic" establishes when there is more than one geographic path for a single entity, the smallest region that is able to ground the whole set is the one that gives them their interpretation. For this heuristic all the geographic coordinates of the non-ambiguous entities are mapped and then a polygon is constructed. The ambiguous entity that is too close or belongs to the formed polygon is considered as the disambiguated entity.

In [13], a probability value is assigned to each possible location in an ambiguous entity. The value is determined by the type of entity, for example, if the entity is a capital or an inhabited place. After that, a set of heuristics is used in order to increment the initial value. These heuristics are: (i) the occurrence of the places, (ii) population statistics and, (iii) geographic terms as city or region.

In [14], Zubizarreta et al. explores different methods of toponym disambiguation based on the frequency of the words and the scores associated to each entity. The recognized geographic paths are structured in a graph form. Then, one filter is used when multiple paths are recognized for a single entity.

Overell et al. in [15] builds a geographic information retrieval system in order to solve the ambiguity problem. They use the co-occurrence of words and Wikipedia. The documents are associated with categories (continent, country, NA). Finally, a set of heuristics is used which allows to search the recognized entities in Wikipedia and determine the disambiguated entity.

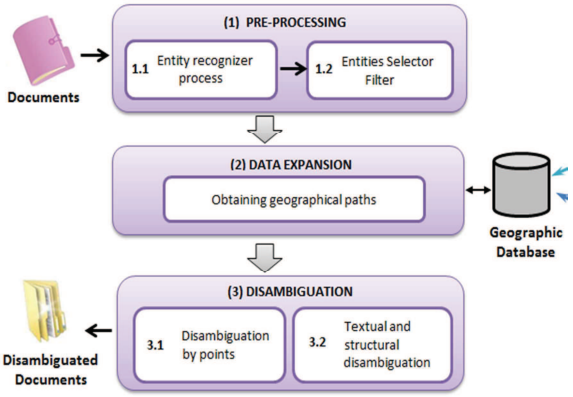
The work in [16], explores the Conceptual Density (CD) technique that is based on Word Sense Disambiguation (WSD) in the geographic context. The CD measures the correlation between the word sense and the document context. The WordNet hierarchies were also used in order to determine the holonymy relations (part-of relations).

The work in [17] explores de toponym disambiguation in geographic information retrieval systems. The focus of this work are the co-occurring toponyms without any ambiguity in the same context of documents, since these toponyms can be regarded as evidences or clues for the process of toponym disambiguation. In order to solve the ambiguity problem the authors proposed an evidence-based approach whose main objective is to calculate the semantic relationships among the geographic references and then use the Dempster-Shafer (D-S). The Dempster-Shafer (D-S) is a theory that combines multiple co-occurring toponyms and allows to determine the right candidate of an ambiguous toponym with multiple evidences.

3 Methodology

The natural language processing is an area of high complexity that has been subject to many researches for several decades. Among the many problems listed in this area, the ambiguity problem is the main focus of this research. In this paper the SpatialCIM methodology is proposed that identifies and geographically locates the documents considering the hierarchical territorial structure of Brazil.

Figure 1: The SpatialCIM Architecture



The proposed methodology architecture is illustrated in Figure 1. This methodology is composed by three main steps (i) pre-processing, (ii) data expansion and (iii) disambiguation. In the (i) pre-processing step, the Named Entities Recognition and Classification in the text (NERC) is carried out with the help of a linguistic tool called Rembrandt [18]. The linguistic tool Rembrandt preforms the NERC task for Portuguese documents. This tool uses Wikipedia as the base of a controlled vocabulary as well as a set of grammatical rules in order to support the extraction of entities. The Rembrandt can recognize entities as person, organization, location, time, among others. After the entities are recognized a filter is applied in order to extract only the geographic

entities. With the selected entities the step of (ii) data expansion begins. In this process all the geographic paths associated to each entity are extracted. A geographic information system (GIS) was used to obtain the geographic paths.

Figure 2: Exemplification of the tree stages of the SpatialCIM

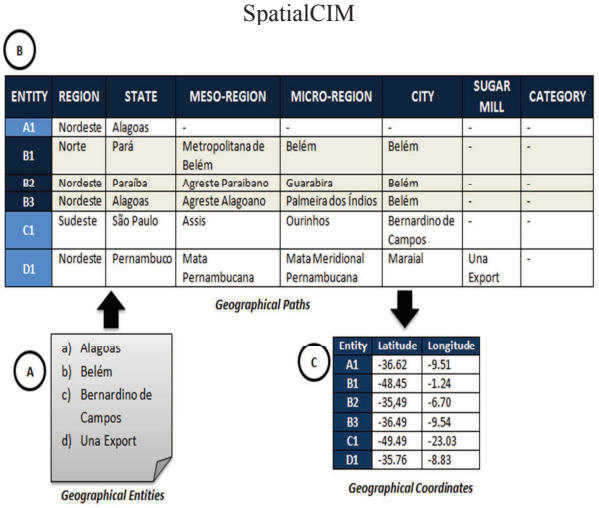
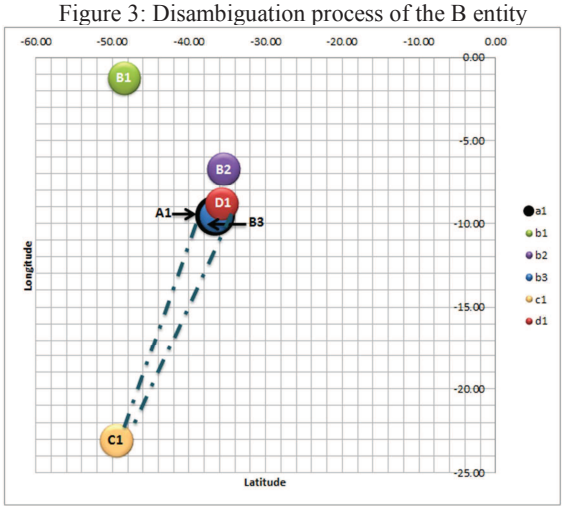


Figure 2 presents an exemplification of the three steps of the SpatialCIM methodology. Figure 2(a) presents a list of the geographic entities extracted from a document that are numbered in order of appearance (a-d). Figure 2(b) illustrates the data expansion performed with the found geographic entities. It can be observed that the recognized entities in Figure 2(a) have associated geographic paths. If there were any ambiguity the paths would look like B1, B2 and B3, that represents the three ambiguous paths for the B entity known as "Belem" as seen in Figure 2(a). For the (iii) disambiguation step it is necessary to obtain all the geographic coordinates of the entities as demonstrated in the Figure 2(c).

The disambiguation process maps the non-ambiguous entities found (A1, C1 and D1) and builds a polygon between them as presented in Figure 3. After the polygon is formed the ambiguous entities (B1, B2 and B3) are mapped as well as the ambiguous entity that belongs to the polygon or the ambiguous entity that is the closest to any point of the polygon is considered as the disambiguated entity. If the ambiguity problem persists a set of heuristics is applied until only one ambiguous entity is selected as the disambiguated entity. Figure 3 illustrates the formed polygon by the A1, C1 and D1 non-ambiguous entities. In order to carry out the disambiguation process the B1, B2 and b3 entities are also mapped. The entity that belongs to the polygon and therefore considered as the disambiguated entity is the B3 entity.

Figure 3: Disambiguation process of the B entity



At the end of the process, the news is marked with their respective associated geographic paths.

4 Experiments and Results

The document collection used in the experiments consists of a set of news focused on the sugar cane culture and provided by Embrapa Agricultural Information[1]. The collection is formed by 237 news documents and it has 593 geographic entities labeled by an expert. Each document is composed by approximately 350 words.

In order to compare the results of the research, the set of labeled news was compared with the results of: (i) Disambiguated geographic paths and the (ii) non-disambiguated geographic paths, in order to determine if the disambiguation process helps with better document localization.

Table 1: Geographic paths recognized and labeled by an expert

News	Region	State	Meso Region	Micro Region	City	Sugar Mill	Category
A	-	-	-	-	-	-	Brazil
	Nordeste	Bahia	-	-	-	-	-
	Sudeste	Minas Gerais	-	-	-	-	-
	Nordeste	Alagoas	-	-	-	-	-
	Nordeste	Ceará	-	-	-	-	-
	Nordeste	Maranhão	-	-	-	-	-
B	-	-	-	-	-	-	Brazil
	-	-	-	-	-	-	Internat.
C	-	-	-	-	-	-	Internat.
	Centro-Oeste	Mato Grosso	-	-	-	-	-
	Centro-Oeste	Goiás	-	-	-	-	-
	Norte	Tocantis	-	-	-	-	-
	Sudeste	São Paulo	Rib. Preto	Rib. Preto	-	-	-
	Sudeste	São Paulo	Met. São Paulo	-	-	-	-

[1] Embrapa: the brazilian agricultural research Corporation (http://www.embrapa.br/english)

Table 1 illustrates the geographic paths related to the sugar mill locations and labeled by the Embrapa expert, Table 2 illustrates the non-disambiguated geographic paths and Table 3 illustrates the disambiguated geographic paths for the "A", "B" and "C" news.

Table 2 Non-disambiguated geographic paths

News	Region	State	Meso Region	Micro Region	City	Sugar Mill	Category
A	-	-	-	-	-	-	Brazil
	Nordeste	Alagoas	-	-	-	-	-
	Nordeste	Bahia	-	-	-	-	-
	Sudeste	Minas Gerais	-	-	-	-	-
	Nordeste	Pernambuco	Met. Recife	Recife	**Recife**	-	-
	Nordeste	**Pernambuco**	-	-	-	-	-
B	-	-	-	-	-	-	Brazil
	Sudeste	São Paulo	Rib. Preto	Franca	**Franca**	-	-
	Sudeste	São Paulo	Rib. Preto	**Franca**	-	-	-
	-	-	-	-	-	-	**Internat.**
C	-	-	-	-	-	-	Brazil
	-	-	-	-	-	-	Internat.
	Centro-Oeste	Goiás	Noroeste Goiano	Rio Verm.	**Goiás**	-	-
	Centro-Oeste	**Goiás**	-	-	-	-	-
	Nordeste	Paraiba	Sert. Paraib.	Catole Rocha	**Mato Grosso**	-	-
	Centro-Oeste	**Mato Grosso**	-	-	-	-	-
	Sudeste	São Paulo	Rib. Preto	Rib. Preto	**Rib. Preto**	-	-
	Sudeste	São Paulo	Rib. Preto	**Rib. Preto**	-	-	-
	Sudeste	São Paulo	**Rib. Preto**	-	-	-	-
	Sudeste	São Paulo	Met. São Paulo	São Paulo	**São Paulo**	-	-
	Sudeste	São Paulo	Met. São Paulo	**São Paulo**	-	-	-
	Sudeste	**São Paulo**	-	-	-	-	-
	Sudeste	Minas Gerais	Zona Mata	Uba	Tocan.	-	-

Table 3 Disambiguated geographic paths

News	Region	State	Meso Region	Micro Region	City	Sugar Mill	Category
A	-	-	-	-	-	-	Brazil
	Nordeste	Alagoas	-	-	-	-	-
	Nordeste	Bahia	-	-	-	-	-
	Sudeste	Minas Gerais	-	-	-	-	-
	Nordeste	Pernambuco	-	-	-	-	-
B	-	-	-	-	-	-	Brazil
	-	-	-	-	-	-	Internat.
C	-	-	-	-	-	-	Brazil
	-	-	-	-	-	-	Internat.
	Sudeste	Minas Gerais	Zona Mata	Uba	Tocan.	-	-
	Centro-Oeste	Goiás	Norest. Goiano	Rio Verm.	Goiás	-	-
	Centro-Oeste	Mato Grosso	-	-	-	-	-
	Sudeste	São Paulo	Rib. Preto	Rib. Preto	-	-	-
	Sudeste	São Paulo	Met. São Paulo	-	-	-	-

In Table 2, the generated ambiguous paths are marked in a different color. In the "A" news the Recife and Pernambuco entities with their ambiguous paths can be observed, because in Brazil the Recife entity is also known with the alternative name of Pernambuco.

In Table 3 the ambiguous path presented in Table 2 was solved. As presented in Table 2 the ambiguity problem referred to the "Pernambuco" and "France" entities was detected. Then, in Table 3 the ambiguity problem was solved and as result the obtained disambiguated geographic paths are similar to the paths marked by the expert as demonstrated in Table 1.

Table 4 Comparison of recall, precision and f-measure for the disambiguated and non-disambiguated process

	Disambiguated process	Non-disambiguated process
Recall	0.6155	0.7116
Precision	0.4185	0.3273
F-measure	0.4982	0.4484

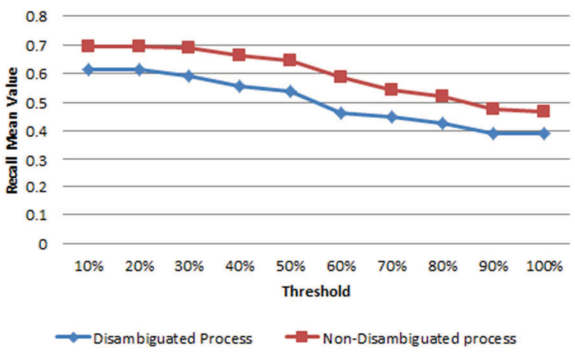
As observed in Table 4, the disambiguated process improves in precision but it decreases the recall measure. This decrease in the recall value is due to the number of recognized entities in each process as observed in Table 5.

Table 5 Comparison of the number of recognized entities and the correct entities with the disambiguated and the non-disambiguated process

	Disambiguated process	Non-disambiguated process
# Recognized entities	872	1289
# Correct recognized entities	365	4422

Notice that in Table 2 there are many ambiguous paths but some of these paths represent the correct paths labeled by the expert as observed on Table 1.

Figure 4: Recall measure for the disambiguated and non-disambiguated process



As explained before and demonstrated in Figure 4 and Figure 5, the disambiguated process has a lower recall value but it improves in the precision value.

Figure 5: Precision measure for the disambiguated and non-disambiguated process

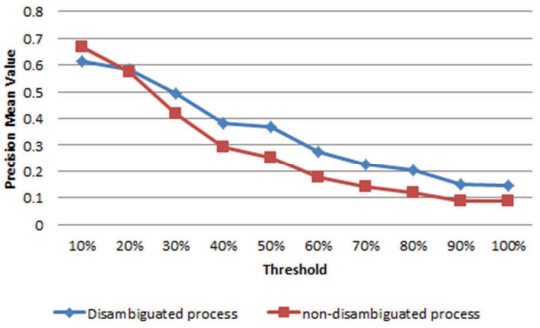
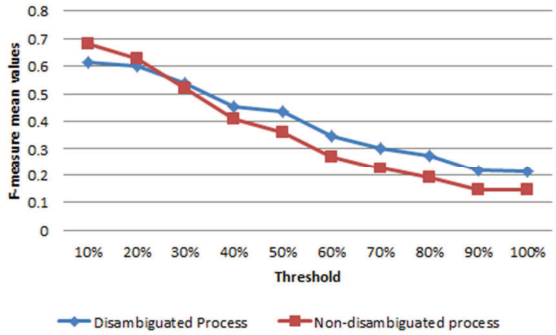


Figure 6: F-measure for the disambiguated and non-disambiguated process



In Figure 6 the f-measure for both processes is presented, demonstrating a better performance for the disambiguated process. The disambiguated process has recognized the geographic paths considering not only the correct number of recognized entities but also considering the total number of recognizes entities.

5 Conclusions

In this paper the process to determine the spatial coverage of the documents focusing on the disambiguation process was presented. The geographic paths labeled by an expert and the obtained disambiguated and non-disambiguated recognized geographic paths were compared.

The experiments demonstrate that even the non-disambiguated geographic paths have recognized a larger number of correct entities, works with a higher number of entities. On the other hand, the disambiguated geographic paths have recognized fewer entities and also the number of correct recognized entities is similar to the number of entities found by the specialist.

The experiments also presented a better performance in the geographic localization of the news when the disambiguation process is used, considering the f-measure.

Acknowledgements We'd like to thank the University of São Paulo, Computer Science Department; Embrapa Agricultural Information and CNPQ by the support.

References

[1] Sanderson, Mark, and Janet Kohler. "Analyzing geographic queries." 27th Annual International ACM SIGIR Conference. Sheffield, UK, 2004.

[2] Mishne, Gilad, and Maarten Rijke. "A Study of Blog Search." Advances in Information Retrieval 3936 (2006): 289-301.

[3] Jones, Rosie, Ahmed Hassan, and Fernando Diaz. "Geographic features in web search retrieval." In Proceeding of the 2nd international workshop on Geographic information retrieval, 57--58. Napa Valley, California, USA: ACM, 2008.

[4] Feldman, Ronen, and James Sanger. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2006.

[5] D'hondt, Joris, Paul-Armand Verhaegen, Joris Vertommen, Dirk Cattrysse, and Joost Duflou. "Topic identification based on document coherence and spectral analysis." Information Sciences 181 (2011): 3783-3797.

[6] Chung-Hong, Lee, Hsin-Chang Yang, and Shih-Hao Wang. "An image annotation approach using location references to enhance geographic knowledge discovery." Expert System Applications 38 (2011): 13792-13802.

[7] Mantratzis, Constantine, Mehmet A. Orgun, and Steve Cassidy. "Separating XHTML content from navigation clutter using DOM-structure." HYPERTEXT 2005, Proceedings of the 16th ACM Conference, 2005: 145-147.

[8] Farazi, Feroz, Vincenzo Maltese, Fausto Giunchiglia, and Alexander Ivanyukovich. "A Faceted Ontology for a Semantic Geo-Catalogue." The Semanic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011 6644 (2011): 169-182.

[9] Buscaldi, Davide, and Bernardo Magnini. "Grounding toponyms in an Italian local news corpus." Proceedings of the 6th Workshop on Geographic Information Retrieval. New York, NY, USA: ACM, 2010. 1--5.

[10] Hill, Linda L., Gail Hodge, and David Smith. "Digital gazetteers: integration into distributed digital library services." Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. New York, NY, USA: ACM, 2002. 427.

[11] Amitay, Einat, Nadav Har'el, Ron Sivan, and Aya Soffer. "Web-a-where: Geotagging Web Content." Proceedings

of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom: ACM, 2004. 273--280.

[12] Leidner, Jochen Lothar. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Boca Raton: Universal Press, 2008.

[13] Li, Yi, Alistair Moffat , Nicola Stokes , and Lawrence Cavedon. "Exploring Probabilistic Toponym Resolution for Geographic Information Retrieval." Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, 2006: 17-22.

[14] Zubizarreta, Ávaro, et al. "A georeferencing multistage method for locating geographic context in web search." Proceeding of the 17th ACM conference on Information and knowledge management. Napa Valley, California, USA: ACM, 2008. 1485--1486.

[15] Overell, Simon, João Magalhães , and Stefan Rüger. "Place disambiguation with co-occurrence models." Conference on Multilingual and Multimodal Information Acess Evluation 2006 Workshop, Working notes (CLEF). 2006.

[16] Buscaldi, Davide, and Paulo Rosso. "A conceptual density-based approach for the disambiguation of toponyms." International Journal of Geographic Information Science, 2008: 301--313.

[17] Wang, Xingguang, Yi Zhang, Min Chen, Xing Lin, Hao Yu, and Yu Liu. "An evidence-based approach for Toponym Disambiguation." The 18th International Conference on Geoinformatics: GIScience in Change. Beijing, China: IEEE, 2010. 1-7.

[18] Cardoso, Nuno. "Rembrandt - reconhecimento de entidades mencionadas baseado em." Encontro do Segundo HAREM, International conference on Computational Processing of the Portugues Language 2008 (PROPOR). Aveiro - Portugal, 2008. 195-211.