The SpatialCIM methodology for spatial document coverage disambiguation and the entity recognition process aided by linguistic techniques

Rosa Nathalie Portugal Vargas [1], Maria Fernanda Moura [2], Eduardo Antonio Speranza [2], Ercilia Rodriguez [2], Solange Oliveira Rezende [1]

[1] University of São Paulo, Computer Science Department
{nathalie, solange}@icmc.usp.br
[2] Embrapa Agricultural Information
{fernanda, speranza, erciliasr}@cnptia.embrapa.br

Abstract. Nowadays it is becoming more usual for users to take into account the geographical localization of the documents in the retrieval information process. However, the conventional retrieval information systems based on key-word matching do not consider which words can represent geographical entities that are spatially related to other entities in the document. This paper presents the SpatialCIM methodology, which is based on three steps: pre-processing, data expansion and disambiguation. In the pre-processing step, the entity recognition process is carried out with the support of the Rembrandt tool. Additionally, a comparison between the performances regarding the discovery of the location entities in the texts of the Rembrandt tool against the use of a controlled vocabulary corresponding to the Brazilian geographic locations are presented. For the comparison a set of geographic labeled news covering the sugar cane culture in the Portuguese language is used. The results showed a F-measure value increase for the Rembrandt tool from 45% in the non-disambiguated process to 0.50 after disambiguation and from 35% to 38% using the controlled vocabulary. Additionally, the results showed the Rembrandt tool has a minimal amplitude difference between precision and recall, although the controlled vocabulary has always the biggest recall values.

Keywords: Named Entity Recognition and Classification, Toponym Resolution, Ambiguity Problem.

1 Introduction

Internet has become a huge online repository of documents, news and several information. This repository is not easily interpretable, thus, requiring tools and techniques to organize, structure and extract interesting information from its documents. For instance, it is necessary to analyze the data structure, similar characteristics and geographic coverage in order to find new trends, problems and solutions for each geographic zone. It is becoming more usual for users to retrieve the information from the context and then consider the geographic location of the document [1]. For exam-

ple, a user searches the context “Main Artificial Intelligence conferences” and then the geographic localization can be specified as “in the United States”. This kind of search is also applied in the search of documents, for example, “Sugar cane production of the last year in Springfield”. However, conventional information retrieval systems based on key-word matching do not consider that words can represent geographic entities which are spatially related to other entities in the document [2]. For instance, for the previous search, the conventional retrieval system will extract all the documents that include in their content the same words used in the search without considering that the word “Springfield” refers to a geographic localization. This process might return a set of irrelevant documents for the user. It be must also considered the ambiguity in the search, for example, “Springfield” is related to three different cities in the United States. In order to determine and extract geographic features from the recognized entities in text it is necessary the development of systems. In this paper we consider all the names that represent a geographic location as geographic entities, such as, the word “Springfield” in the previous example. Some techniques of spatial data mining can be used to allow the geographic feature extraction. Spatial data mining is the process of discovering interesting patterns from spatial databases which were previously unknown but potentially useful [3]; [4].

Much evidence can be considered in the geographic context definition of documents, such as addresses, postal codes, telephone numbers and names of reference points. After the geographic evidence is identified, the possible semantic ambiguities must be checked. Some gazetteers or geographic information systems (GIS) are commonly used to identify the geographic references. A gazetteer is a catalog of locations or places (dictionary of toponyms) which provides a vocabulary of geographic terms along with their respective locations [5]; [6]. The ambiguity can be understood as an expression of the language (word or phrase) which has many different meanings and can be understood in different ways by a receiver. For example, if the city of Venice is mentioned in the text it is necessary to apply a disambiguation process since it can refer to Italian, Colombian or French cities. This ambiguity problem causes noises in the information retrieval process, therefore, the same term can be related to relevant and irrelevant information [7]. The disambiguation process must be applied to solve the ambiguity problems. The disambiguation processes of entities (toponym[1]) are responsible for finding out the spatial location in the text through their standardization in a structured representation as geographic coordinates, database entries, or locations with a geographic ontology [8].

In order to perform the disambiguation process, it is first necessary to identify the geographic entities in the text; this process of identification is known as named entity recognition process (NER). The NER in text is a complex task that aims at locating and classifying atomic elements in the text into predefined categories, such as persons, locations, organizations, expressions of time and money, among others. In fact there are some tools that perform the NER task in text. Such as, the Rembrandt linguistic tool [9] that performs the NER task in Portuguese documents. This tool uses

[1] Toponym is a proper noun designating a place (country, city, continent, etc.)

Wikipedia as the base of a controlled vocabulary as well as set of grammatical rules in order to support the extraction of the entities. A controlled vocabulary is a list of words and phrases used to tag units of information. The Rembrandt tool was developed with the intention of recognizing the entities with a strong connection with geographic places such as countries, cities, rivers, universities, among others.

This paper presents the Spatial Coverage Identification Methodology (SpatialCIM) which allows the identification of the spatial coverage of the text, focusing on news in the Portuguese language. This methodology is based on three steps: pre-processing, data expansion and disambiguation. In the pre-processing step, the Named Entity Recognition and Classification (NERC) process is carried out with the support of the Rembrandt linguistic tool. The performance of the entity recognition process of the documents using the Rembrandt linguistic tool is evaluated against the use of a controlled vocabulary, representing the Brazilian geographic entities.

The set of news used in the experimental set up is related to the sugar cane culture in Brazil. This set of news was manually labeled by a geographic expert. Moreover, the Brazilian sugar mills named in the text and their respective geographic paths are considered as the geographic coverage of documents. The representation of the geographic paths follows the geopolitical hierarchical structure of “Region, State, Meso-Region, Micro-Region, City, Sugar Mill and Category", according to the territorial division of Brazil, as recognized by the Brazilian Institute of Geography and Statistics (IBGE[2]).

This paper is organized as follows: Section 2 covers related work; Section 3 introduces the proposed methodology to determine the spatial coverage of documents; Section 4 presents the preliminary experiments and obtained results, followed by the final considerations in Section 5.

2 Related Work

After recognizing the entities in the text, the following step is to obtain the geographic paths and solve the ambiguity problems in order to establish the correct geographic classification of the documents (or news, in the context of this work). In the literature, many approaches that solve the ambiguity problems are presented. In this paper, we are interested in exploring the (i) toponym resolution and (ii) ontology approach for the disambiguation process.

2.1 Toponym Resolution Approach

The toponym resolution can be defined as the task of setting a reference to a possible ambiguous place name related to its real localization, which is represented in a given context. The proposed approach geographic features are usually obtained with the help of gazetteers or geographic information systems (GIS). For the work present-

[2] http://www.ibge.gov.br/home/

ed here, the interest is to explore the disambiguation method by means of the toponym resolution.

The work in [10] explores different toponym disambiguation methods based on the word frequency and the associated rankings of each entity. The recognized geographic paths represent a graph structure. This structure aims at helping the disambiguation process. When multiple geographical paths recognized for one entity exist, the longest path in the graph is considered and the others paths are discarded. If there is still ambiguity in the paths a democratic disambiguation is used. The method proposed for the authors is based on the co-occurrence of words in the documents and also considers the child nodes of each geographic path. The relevance is based on the frequency of the words from each node in the paths. Consequently, the most important geographic path is selected as the disambiguated path.

In [11], a conceptual density technique (CD) based in the Word Sense Disambiguation (WSD) for the geographical domain is used. The CD is given by the correlation between the word sense and the document context. The WordNet sub-hierarchies are used to determine the holonymy relationships (part-of relation) of the words. With the use of this hierarchy, it is easier to distinguish among the different types of localizations sharing the same name.

In [12], Leidner uses two minimalistic heuristics. The first one uses the linguistic technique called "one-referent-by-discourse". This technique assumes that a mentioned place name in the discourse refers to the same location in the whole discourse. The second minimalistic heuristic aims at solving the ambiguity problem. When more than one geographic path for a single entity exists, the lowest region formed among all the entities is the one that gives its interpretation.

2.2 The Ontology Approach

This approach is based on the use of geographic ontologies as the base for the geographic concepts and their relationships. Geospatial ontologies define classes and individuals to represent geographic regions, their features, and the relations among them [13]; [14].

The geographic features obtained from the geographic ontology are also explored in [15]. These features are used to discover the relationships among documents considering different geographic zones. The classification of the documents is done considering the geographic features to train different algorithms based on supervised and unsupervised machine learning techniques.

In [8], geographic ontology structures, as well as the structure of event ontology of the documents are considered for the disambiguation process. The authors affirm that the event ontology contributes for a better understanding of the spatial coverage of the documents. If two entities participate in the same event, it is a strong indicator of the geographic relationship between them. This indicator helps in the disambiguation process. The geographic ontology used by the authors consists of four levels (earth, country, state and city). The ontology of events considers the following classes: people, organization, and geopolitical localization. After the identification of the entities,

Wikipedia is used in order to establish the relationships between the event ontology and the geographic ontology.

3 The SpatialCIM Methodology

The SpatialCIM methodology allows the identification and localization of news considering the geopolitical hierarchical structure of the Brazilian territory with the support of linguistic techniques for the process of document entity recognition. The proposed methodology is based on three main steps: (1) pre-processing, (2) data expansion and (3) disambiguation. This paper explores the use of the Rembrandt linguistic tool as well as a Brazilian controlled vocabulary in the (1) pre-processing step in order to analyze which technique has a better performance in the news disambiguation process. After the entities are recognized, an entity selector filter is applied to extract only the geographic entities, as observed in the Fig. 1(a).

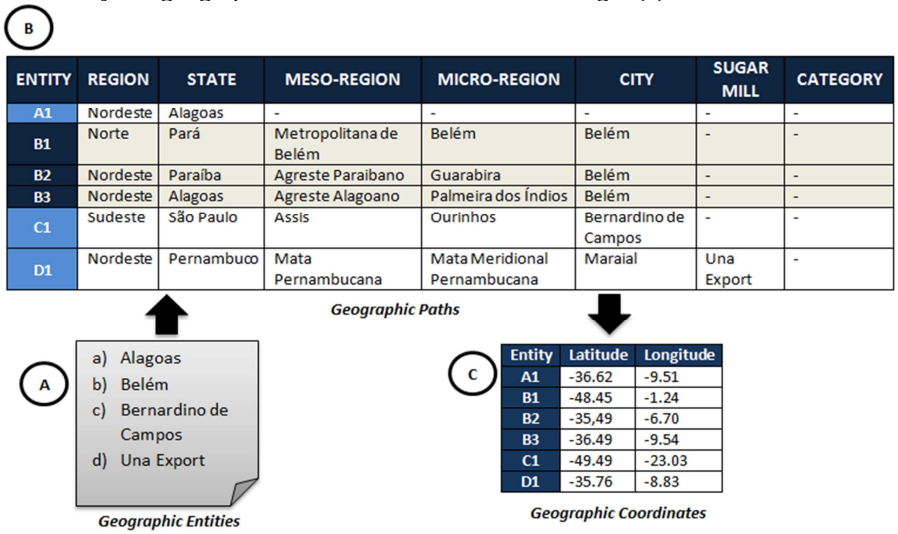

Fig. 1. Three stages of the SpatialCIM

After the geographic entities are extracted, step (2) data expansion, begins. It is necessary to use the spatial data base which contains information about the Brazilian geopolitical hierarchy to extract the geographic paths and the geographic coordinates of the entities. In the SpatialCIM, the spatial data base can be populated using a GIS or a geographic ontology. One example of the geographic paths obtained in this phase is presented in Fig. 1(b), in which the paths obey the hierarchy of Brazil, the Brazilian sugar mill, and the international entities. In Fig. 1(b) the entity column represents the geographic entities found in Fig. 1(a). For instance, for entity (a) “Alagoas” was recognized only as one geographic path marked as A1 was recognized. For entity (b) “Belém” four different geographic paths were recognized, whose paths are marked as B1, B2, B3 and B4 respectively. The same process is repeated for the other ambigu-

ous entities. As observed in Fig. 1(b), the extraction of the geographical paths can present ambiguous paths such as the B1, B2, B3 and B4. This ambiguity problem must be solved in order to allow a more efficient document localization. After all the possible geographic paths are extracted, step (3) disambiguation, begins. This process depends on the system used to populate the spatial data base. If a GIS is used, the applied disambiguation method would be the disambiguation by points. If ontology is selected, the applied disambiguation method would be the textual and structural disambiguation. For the disambiguation by points, the geographic coordinates must be extracted from the spatial data base as illustrated in Fig. 1(c).

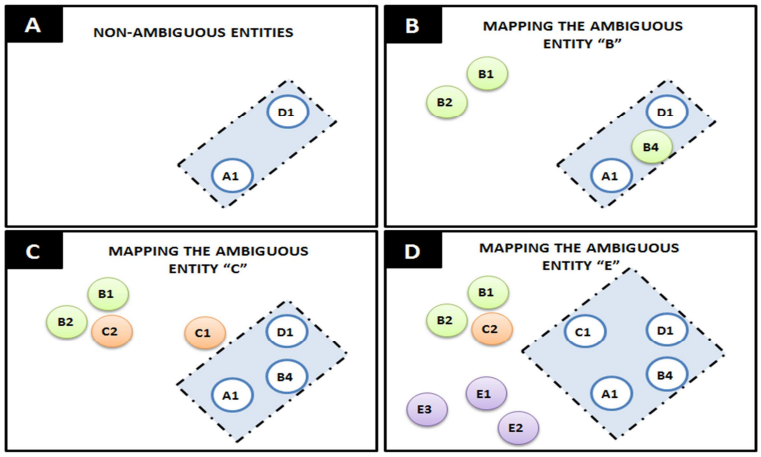

Fig. 2. Disambiguation by the Points Method

In order to execute the process of disambiguation by points, all the non-ambiguous entities (Entity A1 and D1) are mapped in the space and a polygon is formed, as illustrated in Fig. 2(a). Fig. 2(b) presents the disambiguation process for entity "B". This process considers the polygon formed by the non-ambiguous entities showed in Fig. 2(a). Then, the ambiguous entities (B1, B2, B3, and B4) are mapped in the space and the entity that belongs to the polygon or the entity is closest to any points of the polygon is selected. In Fig. 2(b), entity B4 is the disambiguated entity since it belongs to the formed polygon which is marked by a rectangle. This entity is now considered a non-ambiguous entity and a new polygon is formed considering the entities A1, D1 and B4 as demonstrated in Fig. 2(c). After a new polygon is formed the process is repeated until there are no more ambiguous entities as presented in Fig. 2(c) and 2(d).

In the disambiguation process, some problems might occur, such as, the non-existence of non-ambiguous entities at the beginning, or the tie between two ambiguous entities. Next there some cases as well as the solutions proposed are explained.

- If two ambiguous entities belong to the formed polygon, the closest entity to any point of the polygon is considered as correct.
- If the entities do not belong to the polygon and have the same distance to any point of the polygon, then an entity is selected randomly.

- If all the recognized entities are ambiguous, some of the heuristics explained next are used in order to determine the most important entity among all the entities. When the first entity is selected, the disambiguation process continues as previously explained.

The heuristics used to determine the importance of the entities in this paper, considering the experimental set up, are listed next in order of importance:

- In this research the sugar mill entities are more important than other entities.
- As the geographic path considers the hierarchical structure of Brazil, the geographic paths of the entities that are more generic are considered. For example, the entity "São Paulo" is recognized as city and as state. In this case the "state of São Paulo" is selected since it represents a more general path.

At the end of the process, all the documents are marked with their respective associated geographic paths.

4 Experiments and Results

The collection used in the experimental set up is a set of news in the Portuguese language focused on the sugar cane culture. The set of news was marked by an Embrapa[3] expert and considers all the geographic paths associated to each piece of news. The set of documents analyzed in this paper is formed by 237 documents, with 593 recognized entities by the expert, as well as approximately 350 words by document. In the set of news, it is considered that one document can have multiple geographic paths associated.

The main objective of these experiments is to compare the performance of the disambiguation by points using the recognized entities obtained with the help of the Rembrandt linguistic tool and a Brazilian controlled vocabulary. The Rembrandt tool uses linguistics processes in order to determine all the entities in the text. The Rembrandt can automatically recognize entities of the types person, location, time and organization. The Brazilian controlled vocabulary is formed by the intersection of a list with the Brazilian geographic entities and the geographic entities founds in the documents. The process performed by the Rembrandt tool can be computationally expensive since it needs to analyze each phrase and recognize the entities, as well as applying the sense disambiguation. This kind of disambiguation is used when trying to determine the type of entity. For example, the "Paris" entity can be understood as a city, organization, or person entity. If the document is too large, the Rembrandt tool can take considerable time to determine the entities. Considering that linguistic processes are computationally expensive we experimented the use of the controlled vocabulary as a non-linguistic alternative.

To obtain the controlled vocabulary, a list of all the Brazilian geographic entities was considered. One problem with this technique is that the context of the words can-

[3] Embrapa: The Brazilian Agricultural Research Corporation (http://www.embrapa.br /english)

not be recognized; consequently, it can take several types of ambiguities. For example, if the term “São Paulo” is in the text, the controlled vocabulary will recognize this term as a geographic entity without considering the document context. In the document, the term “São Paulo” could be a reference to the São Paulo football team or the state of São Paulo or the São Paulo city.

To achieve the experimental objectives, the geographic paths automatically obtained were compared to the geographic paths marked by an expert. The paths obtained with the entity recognition carried out by the Rembrandt tool and the controlled vocabulary were also compared, in order to determine which technique achieves better geographic paths compared with those obtained by the expert or if the process has no significant differences. To evaluate how much the disambiguation process contributes to better news localization, the resulting geographic paths for each of the news are compared with the geographic paths marked by the expert. Table 1 illustrates the geographic paths of the news marked by the expert. In this case, labels given by the expert are considered as the correct ones for the geospatial classification of the documents. In Table 2, the geographic paths automatically obtained with the entity recognition provided by the Rembrandt linguistic tool can be observed. Observe that two ambiguous geographic paths for the Micro-region of “Catanduva” (rows L1 and L2) and “Jau” (rows L3 and L4) were recognized.

Table 1. Geographic Paths marked by an expert at news 41755

News	*Region*	*State*	*Meso Reg.*	*Micro Reg.*	*City*	*Sugar Mill*	*Categ.*	
41755	Sudeste	São Paulo	-	-	-	-	-	L1
	Sudeste	São Paulo	São Jose Rio Preto	Catanduva	-	-	-	L2
	Sudeste	São Paulo	Ribeirão Preto	Ribeirão Preto	-	-	-	L3
	Sudeste	São Paulo	Pi-racicaba	Piracicaba	-	-	-	L4
	Sudeste	São Paulo	Bauru	Jau	-	-	-	L5

Comparing these results with the results marked by the expert, it can be observed that the “São Paulo”, “Ribeirão Preto” and “Piracicaba” entities (rows L1, L3 and L4 in Table 1) were not recognized. Table 3 illustrates the geographic paths automatically obtained with the entity recognition provided by the controlled vocabulary presented in Table 1.

Table 2. Geographic Paths Generated by SpatialCIM as well as with the Entity Recognition Provided by the Rembrandt tool at news 41755

News	*Region*	*State*	*Meso Reg.*	*Micro Reg.*	*City*	*Sugar Mill*	*Categ.*	
41755	Sudeste	São Paulo	São Jose Rio Preto	Catanduva	Catanduva	-	-	L1
	Sudeste	São Paulo	São Jose Rio Preto	Catanduva	-	-	-	L2
	Sudeste	São Paulo	Bauru	Jau	Jau	-	-	L3
	Sudeste	São Paulo	Bauru	Jau	-	-	-	L4

Table 3 presents ambiguous geographic paths for the “Catanduva” (rows L1 and L2), “Jau” (rows L3 and L4), “Piracicaba” (rows L5, L6 and L7), “Ribeirão Preto” (rows L9, L10 and L11) and “São Paulo” (rows L12, L13 and L14) entities. Obseve that all the entities marked by the expert, in Table 1, were recognized in Table 3. Due to the evidence of ambiguity presented in Table 2 and 3, the disambiguation by points process is applied as detailed in Section 3. Table 2 keeps the L2 and L4 rows and the Table 3 keeps the L2, L4, L5, L8, L9 and L14 rows.

Table 3. Geographic Paths Generated by SpatialCIM as well as with the Entity Recognition Provided by the Controlled Vocabulary at news 41755.

News	*Region*	*State*	*Meso Reg.*	*Micro Reg.*	*City*	*S.M.*	*Cat.*	
41755	Sudeste	São Paulo	São Jose Rio Preto	Catanduva	Catanduva	-	-	L1
	Sudeste	São Paulo	São Jose Rio Preto	Catanduva	-	-	-	L2
	Sudeste	São Paulo	Bauru	Jau	Jau	-	-	L3
	Sudeste	São Paulo	Bauru	Jau	-	-	-	L4
	Sudeste	São Paulo	Piracicaba	Piracicaba	Piracicaba	-	-	L5
	Sudeste	São Paulo	Piracicaba	Piracicaba	-	-	-	L6
	Sudeste	São Paulo	Piracicaba	-	-	-	-	L7
	Nordeste	Pernambuco	Mat. Pernam.	Mata Mer. Pern.	Ribeirão	-	-	L8
	Sudeste	São Paulo	Ribeirão Preto	Ribeirão Preto	Rib. Preto	-	-	L9
	Sudeste	São Paulo	Ribeirão Preto	Ribeirão Preto	-	-	-	L10
	Sudeste	São Paulo	Ribeirão Preto	-	-	-	-	L11
	Sudeste	São Paulo	Met. São Paulo	São Paulo	São Paulo	-	-	L12
	Sudeste	São Paulo	Met. São Paulo	São Paulo	-	-	-	L13
	Sudeste	São Paulo	-	-	-	-	-	L14

As observed, the disambiguation by points and the recognition of entities made with the controlled vocabulary (presented in Table 3) presents a significant number of correct entities (presented in Table 1). The disambiguated results in Table 3 are close to the entities marked by the expert and presented in Table 1.

Table 4. Correct recognized entities with the Rembrandt tool and the controlled vocabulary considering the disambiguated and the non-disambiguated geographic paths

	Disambiguated Geographic paths			*Non-disambiguated Geographic paths*		
	Recall	**Precision**	**F-Measure**	**Recall**	**Precision**	**F-Measure**
Rembrandt tool	0.62	0.42	0.50	0.71	0.33	0.45
Brazilian controlled vocabulary	0.73	0.26	0.38	0.92	0.22	0.35

On the other hand, in Table 4, the general result of the precision, recall and F-measure of the methods for the complete news collection is showed. It can be observed that the Rembrandt tool gives a larger F-measure value of 0.50 over 0.38 obtained by the controlled vocabulary in the disambiguated process. Although, the controlled vocabulary always presents bigger recall values it fails in the precision values. Additionally, the disambiguation process improves the results for the Rembrandt tool as well as controlled vocabulary for both the recall and the precision measures.

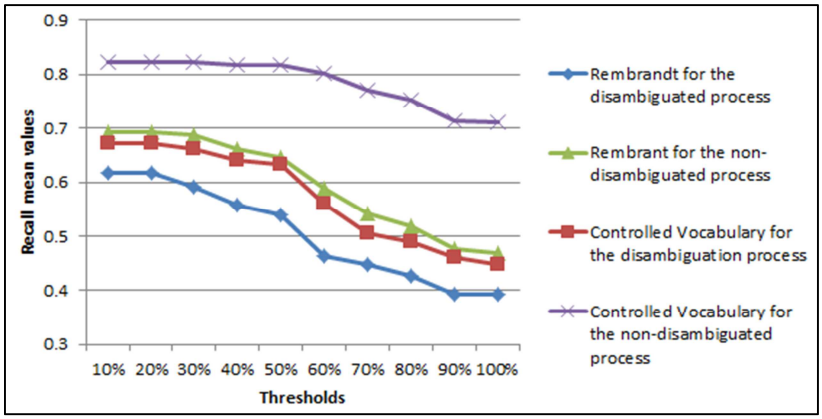

Fig. 3. Recall measure of the Rembrandt tool and the Controlled Vocabulary for the disambiguated and non-disambiguated process.

The presented experiments show the results of the comparison between the Rembrandt tool and the controlled vocabulary for the achievement and disambiguation of the geographic paths.

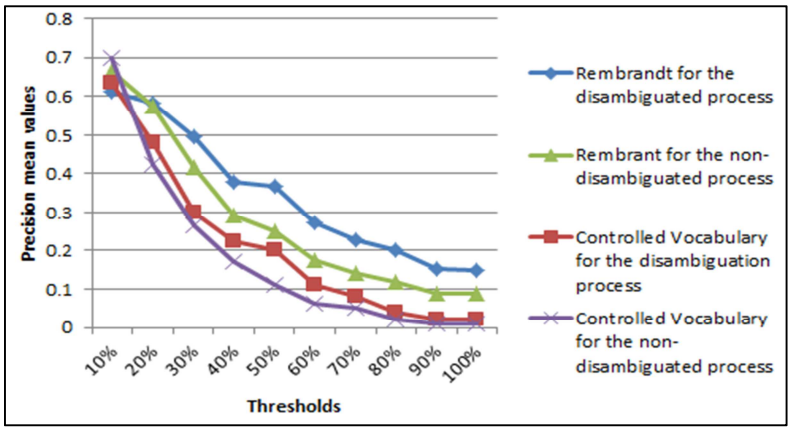

Fig. 4. Precision measure of the Rembrandt tool and the Controlled Vocabulary for the disambiguated and non-disambiguated process.

The results presented in Fig.3 and Fig.4 show the distribution of the recall and precision according to tolerance thresholds. These results also confirm that Rembrandt tool has minimal amplitude difference between recall and precision measure, while recall of the controlled vocabulary is always higher and precision falls. Consequently, the Rembrandt tool is able to obtain a higher F-measure value than the controlled vocabulary.

5 Conclusions

In this paper, the SpatialCIM methodology was detailed which can determine the spatial coverage of documents. Additionally, a comparison between the Rembrandt tool and the use of a controlled vocabulary for the process of recognition of the geographic entities is presented. This comparison was performed in order to evaluate if there was a significant advantage in using the linguistic tool when we also have a controlled vocabulary. For the experimental set up, a set of news in Portuguese about the sugar cane culture, and previously marked by an Embrapa expert was used. Finally, the disambiguation method by points with the support of a GIS and the table of the Brazilian geopolitical division obtained from IBGE was also explored.

The experiments showed a larger F-measure value for the geographic entity recognition process with the use of the Rembrandt tool. It also showed that the disambiguation process actually improves the correct recognition of the entities and it works with a lower number of geographic paths. In future work we intend to analyze the textual and structural disambiguation method with the support of geographic ontologies.

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

References

1. Sanderson, Mark, and Janet Kohler. "Analyzing geographic queries." *27th Annual International ACM SIGIR Conference.* Sheffield, UK, 2004.
2. Jones, Christopher B., Alia I. Abdelmoty, David Finch, Gaihua Fu, and Subodh Vaid. "The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing." *Geographic Information Science Third International Conference, GIScience 2004.* Springer Berlin Heidelberg, 2004. 125-139.
3. Roddick, Jhon F., and Myra Spiliopoulou. "A bibliography of temporal, spatial and spatio-temporal data mining research." *SIGKDD Explor. Newsl.*, 1999: 34-38.
4. Shekhar, Shashi, and Sanjay Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2002.
5. 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.
6. 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.
7. Fuchs, Catherine. L'ambiguïté et la paraphrase: opérations linguistiques, processus cognitifs, traitements automatisés. Centre de publications de l'Université de Caen, 1987.
8. Roberts, Kirk, Cosmin Adrian Bejan, and Sanda M. Harabagiu. "Toponym Disambiguation using events." *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference.* Daytona Beach, Florida: AAAI Press, 2010. 271 -- 276.
9. Cardoso, Nuno. "Rembrandt - reconhecimento de entidades mencionadas baseado em relações de análise detalhada do texto." *Encontro do Segundo HAREM, PROPOR 2008.* Aveiro - Portugal, 2008. 195-211.
10. 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.
11. Buscaldi, Davide, and Paulo Rosso. "A conceptual density-based approach for the disambiguation of toponyms." *International Journal of Geographical Information Science*, 2008: 301--313.
12. Leidner, Jochen Lothar. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Boca Raton: Universal Press, 2008.
13. Jones, Christopher, Alia Abdelmoty, and Gaihua Fu. "Maintaining Ontologies for Geographical Information Retrieval on the Web." In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE - OTM Confederated International Conferences, CoopIS, DOA, and ODBASE*, 934-951. Sicily, Italy: Springer, 2003.
14. Stuckenschmidt, Heiner, and Frank van Harmelen. *Information Sharing on the Semantic Web*. Springer, 2004.
15. Lee, Chung-Hong, Hsin-Chang Yang, and Shih-Hao Wang. "A Location Based Text Mining Approach for Geospatial Data Mining." *Innovative Computing, Information and Control, International Conference*, 2009: 1172 -- 1175.