Geographic Signatures for Semantic Retrieval

David S Batista, Mário J Silva, Francisco M Couto, Bibek Behera
University of Lisbon
Faculty of Sciences, LaSIGE
Lisbon, Portugal
dsbatista@xldb.di.fc.ul.pt

ABSTRACT

Geographic Information Retrieval (GIR) systems rely on the identification and disambiguation of place names in documents to determine the region about which they are relevant. The place names are mapped into geographic concepts and used to assign an encompassing concept (a scope) to each document. However, sometimes a single scope is too restrictive and insufficient for capturing the geographic semantics of a document. We propose as an alternative to abstract the geographic semantics of a document as a geographic signature, which is a list of maximally disambiguated geographic references found in a document. A signature can be used in multiple GIR applications, such as in building a geographic index for a document collection. We perform the disambiguation of the possible geographic meanings using semantic similarity measures.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Design,Experimentation,Measurement

Keywords

Geographic Information Retrieval, NER, Conditional Random Fields, Semantic Similarity

1. INTRODUCTION

In classic Information Retrieval (IR), each document is abstracted as a set of words, ignoring the semantic content and word order (a bag of words). This means that a document can only be retrieved by matching the words it contains and makes it difficult to accurately select documents relevant to a given location. Entering a geographic location name in a classic search engine query might give results that do not relate in any way to that location. To overcame that limitation, the semantic association of locations to documents needs to be extracted. By identifying entities that denote locations in text, using an external knowledge base to associate these entities to concepts and exploring the relations between those concepts, we can enrich the geographic semantics of a document. Words in the document referring to a place can then be mapped into their corresponding geographical information.

Having the semantic geographical information extracted from a given document, and presenting this information in a machine-readable format can be useful for other applications. For instance, a document related to a location can then be retrieved even when the location name is not explicitly given in the query or is not present in the document.

It is common in GIR to capture the geographicity of each document in a single geographic scope inspired in the "one sense per discourse" assumption [4]. However, previous experiments have shown that this approach is too restrictive [1]. We propose in this work to generalize that notion: instead of assigning a document to a single geographic location, we associate a *geographic signature*, which we define as list of disambiguated geographic references found in a document.

The geographic signature is created using text summarization techniques and the knowledge about the relationships among the geographic concepts represented by the words in a document. The geographic concepts and their relationships in the signature can then be used to produce accurate representations of the geographic scopes that characterize each document.

The rest of this paper is structured as follows. Section 2 presents related work. Section 3 shows how the extraction and disambiguation of geographic entities is performed. Section 4 presents initial results of an evaluation of the proposed geographic signatures generations approach. Finally in Section 5 we express our main conclusions and directions for future work.

2. RELATED WORK

The process of grounding names extracted from text to unique places can be tackled in independent phases, and each one has to address different problems. Extracting location names from text is a common natural language processing (NLP) task. To associate semantic knowledge to the extracted names, an external knowledge base must provide the mapping between the geographic names and their corresponding concepts.

We present in this section some of the problems that arise during the process of extracting and associating names to

GIR'10 18-19th Feb. 2010, Zurich, Switzerland

geographic concepts as well as some of the most promising proposed solutions. We also describe the geographic ontology used in this work, which contains the geographic concepts to be included in the signatures.

2.1 Geographic Named Entities Recognition

The task of assigning geographic references to locations can be divided in two main sub-tasks: *geo-parsing* and *geo-coding*.

2.1.1 Geo-Parsing

The *geo-parsing* task is the information extraction step, in which documents are processed to identify geographic references in the text. The extraction of geographic entities from texts is a particular case of the general task of name entity recognition (NER), which identifies expressions in text and classifies them into predefined categories such as persons, events, organizations, places, etc. Two major approaches exist, one based on manually coded rules to define patterns and another based on machine learning [13].

The defined rules, following the grammar of the language in which the document is written, try to find patterns in text that lead to geographic references. The rules usually are supported with dictionaries of place names or gazetteers. Despite these methods achieving good performances, the rules are usually too restrictive and very specific in regard to a type of text.

The machine learning approach is based on extracting features from text that constitute the training data. The features can be surrounding words or properties of the word itself, like capitalization, or frequency of the word in corpus. A probabilistic model is built based on these learning features to identify which can better discriminate when a given word is or not a geographic entity. The model can then be applied to a text identifying potential geographic references.

2.1.2 Geo-Coding

The *geo-coding* task tries to match the names of geographic references identified in a text with related geographic concepts present in ontologies, gazetteers or encyclopedias. During this association phase three types of ambiguity might occur:

1. *referent ambiguity*, when the same name can represent more than one geographic locations; 2. *reference ambiguity*, when the same location is referred to by different names [19]; 3. *referent class ambiguity*, when a name is used to designate locations and other classes of entities, such as persons.

In the first case a geographic concept has to be chosen from a set of geographic concepts. Some heuristics can help on the disambiguation. For instance, one can disambiguate from a set of candidates based on hierarchy levels, such as population, or administrative subdivisions [17]. A large city is more likely of being referred than a small village with the same name. Others simulate natural heuristics employed by humans when reading a text and interpreting geographical references. If the same name is used multiple times in the same text or section, it is assumed that it is always referring to the same location rather than to different locations that share the same name [4]. Another method is to minimize the bounding polygon that contains all candidate referents, using the bounding boxes associated to each concept [10].

To handle the second case one needs to enrich the external knowledge bases with additional data, like alternative names or historical names of places.

The third case is difficult to handle and can be better treated in the *geo-parsing* phase. When recognizing geographic entities in text, the method used should be capable of distinguishing the context in which an entity can represent different classes of entities, places or other. For instance, to identify *Évora*, a city in Portugal, instead of the singer named *Cesária Évora*, it would be required to understand that a text is not about music of Cape Verde.

2.2 Conditional Random Fields

We use Conditional Random Fields (CRF) [9] to recognize geographic named entities in text as a step in our geographic signatures generation process. CRF have been previously used in gene and protein recognition in biomedical abstracts, achieving results current to the state of the art[20].

CRF are a type of discriminative probabilistic model for computing the probability $p(\vec{y}|\vec{x})$ of an output $\vec{y}$ given an input $\vec{x}$, called the observation. This type of model is most used in labeling sequenced structures, such as natural language text. Unlike other models, for instance Hidden Markov Models [16], CRF do not assume strong independence assumptions between the observation variables. A CRF on (X, Y) is specified by a vector $f = (f_1, f_2, ..., f_m)$ of features and a weight vector $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$. The features' values and weights determine the likelihood of each possible value for y_i. From the possible applications regarding CRF, we are interested in exploiting training and classification:

Training: Given a set of training data (x^k, y^k) find the parameters λ of the CRF that will maximize the likelihood with the training data;

Classification: For a given CRF, with parameters λ and an input sequence $\vec{x}$, find the most likely label y such that $y = argmax_y\ p_\lambda(\vec{y}|\vec{x})$;

The training phase generates a model, based on the labels $\vec{y}$, given to the input data $\vec{x}$. The model contains features which can then be used in the classification phase to assign labels $\vec{y}$ to a given input $\vec{x}$. Having training data, texts where the geographic locations are annotated, a CRF model can be generated, using the annotated locations to generate features. The CRF model can then be applied to other texts to extract geographic references.

2.3 Geographic Knowledge Representation

A representation of places from the real world in a structured way can be used by different tasks in spatial search engines, like geographic text mining or query interpretation and reformulation. Usually this knowledge is supported by an ontology, where information regarding places' names, types, spatial footprints and relationships among them is stored.

The geographical type (e.g. city, state, village) and the spatial footprints (bounding boxes, geographic delimitations, centroid coordinates) are used in the extraction of relationships between places. Additional information should also be incorporated, such as population, which can be used to disambiguated place names (e.g. places with higher populations counts or economic activity have a higher probability of being mentioned in documents). The geographic ontology can include historical data, such as alternative and historical place names.

Table 1: Characterization of Geo-Net-PT02

Feature Type	N° Features	(%)
Postal Code	187 014	48.44
Street Segments	146 422	37.93
Settlement	44 386	11.50
Civil Parishes	42 60	0.93
Zone	3 594	0.08
Municipality	308	0.01
NUT	40	0.01
Districts	18	0.00
Province	11	0.00
Island	11	0.00
Region	2	0.00
Country	1	0.00
Total	386 067	100.00

(a) Statistics of the Administrative Domain

Feature Type	N° Features	(%)
Stream	2 421	42.65
Beach	588	9.83
Museum	507	8.93
Archaeological Site	414	7.29
Hotel	381	6.71
Natural Region	304	5.36
Castle	256	4.51
Spring	220	3.88
Historic Hamlet	217	3.82
Reservoir	90	1.59
Touristic Resource	84	1.48
Other	224	3.95
Total	5 676	100.00

(b) Statistics of the Physical Domain

Domain	*part-of*	*adjacent-to*
Administrative	386 431	33 051
Physical	389	2 404
Inter-Domain	2 752	0

(c) Relationships in Geo-Net-PT02

We are developing and evaluating our geographic signatures generation software in an environment composed of Portuguese texts and a geographic ontology of Portugal, Geo-Net-PT02 [12]. This ontology contains about 400,000 administrative and physical features logically divided into two information domains (see statistics organized by feature type in Table 1a and 1b).

For each domain a set of relationships between the different feature types is defined. For instance, in the administrative domain, features of the type civil parish have a *part-of* relationship with municipality features, which in turn are *part-of* districts. Another type of relationship is *adjacent-to*. A district can be *adjacent-to* other districts and two municipalities can be *adjacent-to* to one another while being *part-of* a district. In the physical domain, the two types of relations are also defined for different feature types. There are also relationships between the administrative and physical domain. Table 1c presents the statistics of relationships within each domain and between the administrative and physical domains.

Table 2: Ambiguity in Geo-Net-PT02

Names	Administrative	Physical
N° Names	77 748	5 209
Ambiguous	19 647 (25%)	329 (6%)
Non-Ambiguous	58 101 (75%)	4 880 (94%)

(a) Referent ambiguity in Geo-Net-PT02 names

Feature Type	Total N° Features	N° Features with a non unique name
Street	91 310	58 770 (64.36%)
Travessa	18 150	10 613 (58.47%)
Town square	7 284	4 095 (56.22%)
Avenue	3 630	1 905 (52.48%)

(b) The most ambiguous feature types in Geo-Net-PT02

The ambiguity present in features' names is quantified in Table 2a, both for the physical and administrative domains. An ambiguous name is one which is used as the name of more than one feature in Geo-Net-PT02. More than one quarter of the names present in the administrative domain is ambiguous.

Table 2b shows the four most ambiguously named feature types, which are all street segments. More than 50% of the features belonging to each of those feature types share the name with some other geographic feature.

3. GEOGRAPHIC SIGNATURES

A geographic signature describes extracted geographic references from a text. The geographic references are disambiguated when possible and associated to matching geographic concepts in an ontology. Other information, such as coordinates or bounding boxes for each disambiguated term, is also included in the signatures.

3.1 Geographic Entities Extraction

Some of the extracted geographic references from a document may be mapped into different geographic concepts in an ontology, since the text found is ambiguous. Extracting as much information as possible about the geographic references in the text facilitates the process of disambiguation. Considering Geo-Net-PT02 as an external knowledge base, it helps if the references are extracted together with their feature type name. For instance, if the geographic reference *"Avenida da Liberdade"* is in a text, and only *"Liberdade"* is identified as a geographic place name, we will have to disambiguate from up to 486 different geographic concepts in Geo-Net-PT02. If we extract *"Avenida da Liberdade"* and match all the geographic concepts with name *Liberdade* and feature type *Avenida* the number of returned concepts is reduced to 69.

3.2 Disambiguation Process

The disambiguation process aims at selecting the concept that better describes the semantics of the geographic reference in a document. We propose to use semantic similarity measures [5] for this propose. Given two ontology concepts, semantic similarity measures (SSM) return a nu-

merical value reflecting the closeness in meaning between them. Our proposal for disambiguating an extracted geographical reference in a document is to select the concepts that maximize the semantic similarity within the geographic signature to return, i.e. the concepts closest to the concepts also referenced in the document.

Semantic similarity has been successfully applied in several application domains, such as biomedical [15], WordNet [21] and also in GIS [7]. Several approaches are available to quantify the semantic similarity between concepts of an ontology represented as a directed acyclic graph (DAG), such as Geo-Net-PT02. One technique commonly used in these approaches is Information Content (IC), which gives a measure of how specific and informative a term is [18].

3.2.1 Information Content

The IC of a concept c can be quantified as the negative log likelihood,

$$-\log p(c)$$

where $p(c)$ is the probability of occurrence of c in a specific corpus. The concept of IC is cumulative, that is, the IC of a concept c depends on its descendants in its subtree. As Geo-Net-PT02 is represented as an inverted tree DAG, this would mean that, as we descend the tree, the probability of a concept decreases and hence its IC increases. If there were a single ancestor (or root at the top) of the DAG, it would have a probability of 1 or an IC of 0. The probability of occurrence is normally estimated by the frequency of annotation of the concept. For example, $p(c)$ can be calculated through the number of occurrences of c in web pages. However, referent ambiguity is present, the same terms can refer to different concepts. For example, *Lisboa* can refer to the concept representing the city of *Lisbon* or just a street in other city. Hence, counting the frequency of a concept in the web may cause inconsistencies in IC estimation. This problem may be smoothed by assuming that the IC of a set of geospatial concepts adjacent to each other should follow a normal distribution. In other words, if we choose the cities in the path between *Lisboa* and *Porto* their IC should approximate to a curve without sudden discontinuities.

3.2.2 GeoSpace

We do not usually want to compare two geospatial concepts by the information they share. We want to check how much space and social features they share instead. We propose an alternative to calculate $p(c)$ by measuring the geographical content described by a concept.

We first define $geospace(c)$ as the geographical content that the concept c and its descendants refer to:

$$geospace(c) = \bigcup_{d \leq c} geospace(d)$$

where $d \leq c$ means that d is descendant of c, or c itself.

We assume that there is a concept *root* such that for any concept c of the ontology we have $c \leq root$. Thus, $geospace(root)$ defines the entire geospace modelled by the ontology, for example a country, such as *Portugal*, or the *Earth*.

We can now calculate the value of a given spatial or social feature for a given *geospace*. As main features we propose:

- *area*: the number of spatial area units
- *specificity*: number of concepts that share a $geospace(c)$
- *population*: number of people living in a $geospace(c)$

The *area* and $specificity$ features are only dependent of physical dimensions and of the ontology structure, respectively. The *population* feature can be obtained from demographic data. Depending on the setting, other features *geospace* that may be derived from other demographic indicators rather than *population*, such as: Education; Health; Housing; Human settlements; Income and Economic Activity; Literacy; Unemployment; and Water Supply and Sanitation.

Therefore we estimate $p(c)$ as follows:

$$p(c) = \sum_{i=1} \lambda_i \frac{f_i(geospace(c))}{f_i(geospace(root))}$$

where f_i can be any of the features described above: *area*, $specificity$ and *population* and $f_i(geospace(c))$ is the geographic content of a feature c measured by feature f_i.

3.2.3 Similarity Measures

Assuming that the information shared by two concepts is proportional to their semantic similarity, the concept of IC can be applied to the common ancestors of two terms, to quantify the amount of information they share, and thus measure their semantic similarity.

The most popular semantic similarity measures based on IC have been proposed by Resnik [18], Lin [11], and Jiang and Conrath (JC) [8]. JC's measure relates the IC of the most informative common ancestor (MICA) to the IC of the terms being compared:

$$SSM_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA})$$

where c_{MICA} represents the most informative common ancestor of c_1 and c_2.

3.2.4 Example

For example, having extracted the terms *Lisboa* and *Santa Catarina* from a document, the following associations to concepts in Geo-Net-PT02 can be made:

- *Lisboa* is a municipality (ID_1)
- *Lisboa* is a place in the municipality of Monção (ID_2)
- *Santa Catarina* is a civil parish in the municipality Lisboa (ID_3)
- *Santa Catarina* is a street in the municipality of Porto (ID_4)

Calculating the semantic similarity for each pair of concepts with names (*Lisboa*, *Santa Catarina*), we obtain:

$$SSM\ (ID_1, ID_3) = 0.58$$
$$SSM\ (ID_1, ID_4) = 0.06$$
$$SSM\ (ID_2, ID_3) = 0.06$$
$$SSM\ (ID_2, ID_4) = 0.14$$

For this example, the pair *Lisboa*(ID_1) and *Santa Catarina* (ID_3) have the highest value, meaning that those are most geographically related. Thus, the geographic signature of the document would be composed by ID_1 and ID_3.

3.3 Geographic Signatures Algorithm

The generation of geographic signatures for a given document is divided in 3 phases:

1) **Geo-Parsing**: a trained CRF using pre-annotated texts and dictionaries of location names as learning features, extracts names of geographic entities from texts.

2) **Geo-Coding**: the disambiguation process is made in two steps; first there is an elimination of features that commonly receive their names after other named entities, street segments in particular. The same name can also be used for different geographic concepts, with different feature types. In order to reduce the number of concepts we apply the following heuristics:

2.1) **Elimination of street segments concepts, there are two situations:**

I) A geographic reference is extracted together with a feature type: all the geographic concepts that match the name are used. For instance, having extracted the name *"Avenida da Liberdade"*, *"Avenida"* being a feature type, all the geographic concepts whose type is *"Avenida"* and name is *"Liberdade"* are taken into consideration.

II) A geographic reference is extracted without any feature type: if the associated concepts are only street segments, then none is taken into consideration. For instance, looking for concepts that match the name *"Brasil"* in Geo-Net-PT02, we have 83 concepts referring only to street segments. We assume that if there is a reference in a text to such concepts the feature type is explicitly mentioned with the name reference, i.e *"Avenida do Brasil"*, and not just *"Brasil"*.

Concepts whose feature types are not street segments are all taken into consideration. For the name *"Beja"* 20 concepts are returned: 1 District, 1 Municipality, 3 Civil Parishes and 15 Street Segments. In this case, only the first four, corresponding to higher level concepts are used.

2.2) **Disambiguation**: the geographic SSM is applied to pairs of concepts to select which pair best relates the two terms. The extraction of geographic references from text includes the position of occurrence in the text. An SSM function is applied to every pair of concepts, following the order of occurrence in the text.

For instance, considering the following sentence: *"...he went through Avenida da República to Marquês de Pombal, there he took the subway to Rossio..."*, and having extracted the geographic references: *"Avenida da República"*, *"Marquês de Pombal"* and *"Rossio"* and assuming that we have a set of geographic concepts for each reference, we first apply an SSM function to all possible pairs of concepts for *"Avenida de República"* and *"Marquês de Pombal"*, choosing the pair with the highest value. Next we apply again a SSM function, fixing the chosen concept for *"Marquês de Pombal"* and using all the possible concepts for *"Rossio"*, we choose then the concept for *"Rossio"* that yielded the highest value. In this process, every extracted geographic reference is grounded to one geographic concept only.

3) **Signature Generation**: after the extraction and disambiguation, a geographic signature is generated. It presents all the extracted geographic references in the text, associated to geographic concepts in Geo-Net-PT02. If there are any relationships between the geographic concepts, those are also indicated. The signature is then formatted in RDF for sharing with GIR applications.

4. EVALUATION

In this section we describe how the CRF model was trained, the data sets used, and the results of an experiment using the articles from the Portuguese Wikipedia about the 18 districts of Portugal.

4.1 Geographic References Extraction

We trained a CRF model using Minorthird's implementation of CRF [3]. To train the model, we used the golden collections (GC) from HAREM, an evaluation contest for named entity recognizers in Portugese that had three editions in the past years [14]. The GC of each event contains 10 different types of manually-tagged references in the text collection. To create a training set for the CRF classifier, the GC of the 2005 and 2006 editions of HAREM have been filtered to eliminate all tags unrelated to locations, so that only PLACE tags would be present. Table 3 describes the collections according to occurrences of PLACE tags and text size.

Table 3: Statistics for the PLACE entities in the GC of the 3 HAREM editions.

Properties	2005	2006	2008
Document Size	731 Kb	512 Kb	1098 Kb
Unique PLACE names	488	371	612
Total PLACE names	1099	759	1200

Regarding the selected features, we chose to turn off those that are English dependent, given that we were processing texts in Portuguese only. We added additional features, apart from the manually-tagged places and the features generated by Minorthird, to the CRF model, from dictionary names and in the from of Hearst patterns [6].

Minorthird provides features that describe patterns. For instance, *charTypePattern.9+* describes that the token is composed by numbers only, and *charTypePattern.X+x+* describes a capitalized word. Another possible feature is a lower-case version the word itself, for instances: *eq.lc.avenida* or *eq.lc.aeroporto*. All these features were selected for the neighbouring window [-3,3].

Words which correspond to names of feature types in Geo-Net-PT02 were annotated generating the CRF feature *isFeatureType*. Names of Portuguese districts, municipalities and civil parishes also generated the CRF feature *isGeoName*.

A list of adjectives and verbs that frequently occur before place names were used to generate the CRF feature *isLocalPrefix*, and a list of prepositions generating the CRF feature *isPreposition* [2].

These CRF features present in the training texts as labels were used to build Hearst patterns. The patterns are comprised by three labels and aim at identifying expressions, which can occur close to place names. For example, a

isLocalPrefix label followed by *isPreposition* label followed by *GeoName* label, or a *isFeatureType* label followed by a *isPreposition* label followed by a *GeoName* label. Every expression that matches the defined patterns was annotated as *localPrefixes*.

These annotations were used to generate additional features to the CRF model. The trainning data contains therefore more information than just PLACE tags, that we used as labels for the CRF initial training phase. Table 4 presents the obtained precision, recall, and F-1 score along with results of other systems that participated in the HAREM II evaluation contest for the PLACE tag only track [14].

Table 4: Results for PLACE tag in HAREM II

System	Precision	Recall	F-1
REMBRANDT	0.56	0.73	0.63
SEIGeo	0.71	0.51	0.59
Minorthird	0.69	0.47	0.56
SeRELeP	0.22	0.79	0.34

We evaluated the geographic references extraction performance using a set of pages from the Portuguese Wikipedia describing the 18 districts of Portugal, in which the geographic locations and associated concepts in Geo-Net-PT02 have been manually annotated. Table 5 shows the number of geographic entities annotated in each page, and the precision/recall of the geographic names extraction process. The results vary according to the number of geographic references annotated for each page. The pages of the most populated and economically active districts have more text, and therefore more geographic references. The CRF model that we produced is comparable in terms of F-1 to the best models that participated in HAREM II but has a low recall. We conjecture that this may be caused by having the CRF trained with unsufficient, data which causes the model to overfit to the training data.

4.2 Disambiguation using Geo-Net-PT02

We implemented the computation of the semantic measure introduced in this paper with Geo-Net-PT02. To calculate the IC of each Geo-Net-PT02 concept we used the Google N-Grams[1] corpus.

Given a list of extracted geographical references found in a document, the disambiguation process calculated the semantic similarity between all the associated concepts in Geo-Net-PT02 and returned the list of concepts, one for each reference, that maximized the semantic similarity between them.

We manually evaluated how many of these entities were correctly disambiguated according to the algorithm described in Section 3.3. Table 6 shows the percentage of correctly extracted and the percentage of correctly disambiguated entities.

The results for the disambiguation process are satisfactory, although some analyzed articles show lower percentage of correctly disambiguated entities. One problem tends to be with incorrectly disambiguated names caused by the lack of a corresponding concept that should have been matched.

[1]`http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html`

Table 5: Results of extraction on Wikipedia articles

Page of	Entities	Precision	Recall	F-1
Aveiro	22	0,80	0,58	0,67
Beja	24	0,69	0,37	0,48
Braga	190	0,37	0,51	0,43
Bragança	11	0,56	0,39	0,46
Castelo Branco	23	0,71	0,46	0,56
Coimbra	85	0,52	0,38	0,44
Évora	11	0,90	0,37	0,52
Faro	58	0,68	0,53	0,60
Guarda	46	0,60	0,48	0,53
Leiria	98	0,70	0,44	0,54
Lisboa	225	0,66	0,50	0,57
Portalegre	79	0,41	0,56	0,48
Porto	101	0,40	0,53	0,45
Santarém	22	0,83	0,42	0,55
Setúbal	38	0,73	0,53	0,62
Viana do Castelo	12	0,84	0,48	0,62
Vila Real	51	0,52	0,62	0,57
Viseu	80	0,46	0,59	0,52

For instance, Spain and names of cities of Spain in the descriptions of the borders of districts are extracted from the text. Some of those names also correspond to small and distant locations in Portugal, which obtain the highest similarity.

Other problems are the referent class ambiguities, that is, entities extracted as potential geographic references that in the text refer to another class of entity, mostly person names or historic figures. In Wikipedia pages, these usually occurr in the paragraph that describes the history of the district, and some of these extracted names are matched to geographic concepts in the ontology.

The semantic similarity measures were applied based on the order of the extracted geographic references in the texts. When a person's name is erroneously extracted as a geographic reference that matches concepts in Geo-Net-PT02, the disambiguation approach based on semantic similarity will give false results that can propagate to the rest of the disambiguation process. An alternative is to perform a more complex disambiguation, comparing all the names in a sentence of vicinity of each concept.

5. CONCLUSIONS AND FUTURE WORK

We presented an initial version of a geographic signatures generator. We used for the *geo-parsing* phase a machine learning approach based on CRF. The recall of the trained CRF model is still relatively low. We believe that a better model can be generated through the tuning of the generated and selected features during the training phase. Nevertheless, the lack of large Portuguese labelled corpus for training the CRF is likely the biggest limitation.

Another aspect concerning the CRF, is it's ability to capture the feature types associated to a given entity, which may, in the best case, completely eliminate the need of disambiguation. In some cases the trained CRF model fails to capture evidences close to the geographic reference that explicitly refer the geographic feature type (i.e.: municipality, district, etc). It should be possible to generate a better CRF

Table 6: Results for the extracted entities

Page of	Correctly Extracted	Correctly Disambiguated
Aveiro	100%	70%
Beja	88%	87%
Braga	71%	67%
Bragança	100%	75%
Castelo Branco	38%	54%
Coimbra	70%	82%
Évora	100%	100%
Faro	80%	68%
Guarda	93%	76%
Leiria	90%	85%
Lisboa	96%	92%
Portalegre	90%	68%
Porto	87%	68%
Santarém	100%	81%
Setúbal	81%	70%
Viana do Castelo	100%	62%
Vila Real	77%	83%
Viseu	92%	89%

model capable of handling the detected limitations.

For *geo-parsing*, we based our work in a geographic ontology covering the territory of Portugal and used semantic similarity measures to desambiguate. We defined an Information Content value based on the probability of the name of each geographic concept in a given corpus. In the future, we intend to use different semantic measures and calculate the Information Content using the proposed definitions of *geospace*.

We plan to generate geographic signatures for each document that is part of WPT05[2], a crawl of the Portuguese web and later evaluate the effectiveness of geographic signatures in GIR. We also plan to build a similar prototype to handle documents in English.

6. ACKNOWLEDGEMENTS

This work was supported by FCT (Portugal), through the project PTDC/EIA/73614/2006 (GREASE-II) and the Multiannual Funding Programme.

7. REFERENCES

[1] Nuno Cardoso, Mário J. Silva, and Bruno Martins. The University of Lisbon at CLEF 2006 Ad-Hoc Task. In Carol Peters et. al, editor, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 51–56, Berlin / Heidelberg, 2007. Springer. Revised Selected papers.

[2] Marcirio Silveira Chaves. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*, pages 231–245, 2008.

[2] `http://xldb.di.fc.ul.pt/wiki/WPT_05_in_English`

[3] William W. Cohen. Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, http://minorthird.sourceforge.net. 2004.

[4] William A. Gale, Kenneth W. Church, and David Yarowsky. One Sense per Discourse. In *HLT '91: Proceedings of the Workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[5] R. L. Goldstone and J. Son. *Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press, New York, 2005.

[6] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *In Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[7] Krzysztof Janowicz, Martin Raubal, Angela Schwering, and Werner Kuhn. Semantic Similarity Measurement and Geospatial Applications. *T. GIS*, 12(6):651–659, 2008.

[8] David W. Conrath Jay J. Jiang. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Computing Research Repository*, cmp-lg/9709008, 1997.

[9] John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[10] Jochen L. Leidner, Gail Sinclair, Bonnie Webber, and Edinburgh Eh Lw. Grounding Spatial Named Entities for Information Extraction and Question Answering. In *In Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL'03*, pages 31–38.

[11] Dekang Lin. An Information-Theoretic Definition of Similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[12] Francisco J. Lopez-Pellicer, Marcirio Chaves, Catarina Rodrigues, and Mário J. Silva. Geographic Ontologies Production in GREASE-II. Technical Report TR 09-18, University of Lisbon, Faculty of Sciences, LASIGE, November 2009.

[13] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[14] Cristina Mota and Diana Santos. *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM*. 2008.

[15] Catia Pesquita. Improving Semantic Similarity for Proteins based on the Gene Ontology. Master's thesis, University of Lisbon, Faculty of Sciences, December 2007.

[16] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, pages

257–286, 1989.

[17] Erik Rauch, Michael Bukatin, and Kenneth Baker. A Confidence-Based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 50–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[18] Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI*, pages 448–453, 1995.

[19] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *CoRR*, cs.CL/0306050, 2003.

[20] Burr Settles. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *JNLPBA '04: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[21] Dongqiang Yang and David M. W. Powers. Measuring Semantic Similarity in the Taxonomy of Wordnet. In *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science*, pages 315–322, Darlinghurst, Australia, Australia, 2005. Australian Computer Society, Inc.