The XLDB Group at GeoCLEF 2005

Nuno Cardoso, Bruno Martins, Marcirio Silveira Chaves, Leonardo Andrade and Mário J. Silva
Grupo XLDB - Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
{ncardoso, bmartins, mchaves, leonardo, mjs} at xldb.di.fc.ul.pt

Abstract

This paper describes our participation at the GeoCLEF 2005 task. We detail the main software components of our Geo-IR system, its adaptation for the participation at GeoCLEF and discuss the obtained results. The software architecture includes a geographic knowledge base, a text mining tool for geo-referencing documents, and a geo-ranking component to re-rank the results of a standard IR index according to geo-scopes. Evaluation shows that ranking with geographic scopes is heavily dependent on the information loaded in the knowledge base and on the ranking algorithm involved, requiring more than the correct assignment of a geo-scope to each document.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Experimentation, Design

Keywords

Geographic Information Retrieval, Text Mining, Evaluation, CLEF, GeoCLEF

1 Introduction

Over the past two years, the XLDB Group has been developing and operated tumba!, a search engine for the Portuguese community, publicly available at `http://www.tumba.pt` [18]. Tumba! is a testbed for several data management, information retrieval and data mining research efforts. We are currently extending it to handle geographic searches, under the GREASE (Geographical REAsoning for Search Engines) project.

GREASE researches methods, algorithms and software architecture for geographical information retrieval (Geo-IR) from the web [6, 19]. Some of the specific challenges are: 1) building geographical ontologies to assist Geo-IR; 2) extracting geographical references from text; 3) assigning geographical scopes to documents; 4) ranking documents according to geographical relevance.

GeoTumba, a location-aware search engine handling *concept@location* queries, is a prototype system developed in the context of GREASE.

Last year, the XLDB Group made its debut participation in CLEF, using tumba! at the monolingual ad hoc Portuguese retrieval task [2]. This year, along with a second participation in the ad hoc task, we also entered the GeoCLEF task, to evaluate the work done so far on GeoTumba. Our intent was to obtain results that could provide interesting insights on the validity of our approaches. In order to build a system configuration that would enable us to generate the runs, we made significant adaptations to GeoTumba,

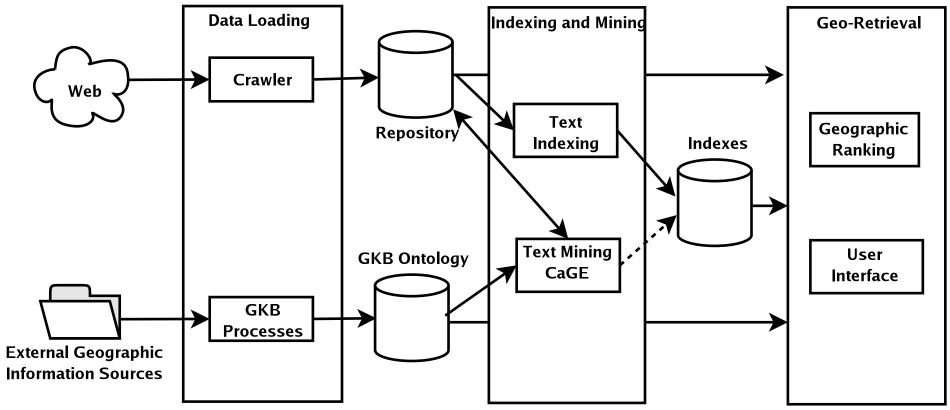

Figure 1: The Geographic IR architecture

including using global geographic information instead of just focusing on the Portuguese territory, and replacing the ranking component (which is still under development) by a simpler scheme.

The rest of the paper is organized as follows: Section 2 describes GeoTumba and the software configuration that was assembled for our participation at GeoCLEF. Section 3 outlines our evaluation goals and the submitted runs. Section 4 presents an analysis on the obtained results, and finally, Section 5 draws conclusions and directions for future work.

2 The Geographic IR system

Figure 1 shows the architecture of the current Geo-IR system prototype. Information is processed in three phases:

Data loading: web pages are harvested into a repository by a crawling module. The geographic knowledge of GeoTumba is collected into GKB (Geographic Knowledge Base) [3, 4]. GKB can be queried interactively to retrieve data about a geographic name or a relationship about two geographic features. It can also be used to create geographic ontologies.

Indexing and Mining: the geographic ontology is used by CaGE, a text mining module for recognizing geographical references and assigning documents with a corresponding geo-scope [10, 11]. Once the scopes are assigned to documents, we create indexes for a fast retrieval of web documents. The indexing software of tumba! is being enhanced for indexing the geographic scopes information. Term indexes are created from the repository data to handle the *concept* part of the queries, while the *location* part is used as a key for fast access to documents through the scopes indexes.

Geo-Retrieval: in the last phase, the indexes and repositories previously created are accessed to generate the result sets that match users' queries, ranked according to geographic criteria.

In the rest of this Section, we describe the main modules, namely GKB and CaGE. The Section ends with a description of the software configuration that we assembled to create the runs submitted to GeoCLEF.

2.1 GKB

Geographic knowledge is collected into a common knowledge repository, the GKB. Its purpose is both to provide a common place for integrating data from multiple external sources under a common schema, and to support mechanisms for exporting geographic knowledge to be used by other components.

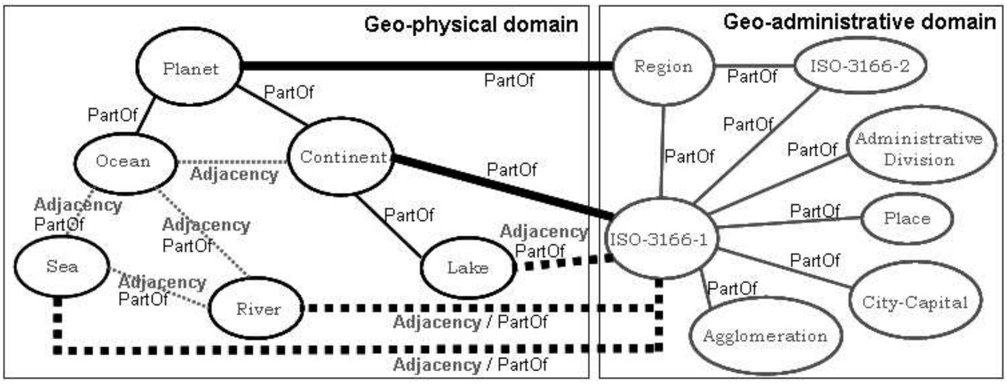

Figure 2: Feature types and their relationships in the world ontology of GKB

The geographical information in GKB includes names for places and other geographical features, information types (e.g. city, street, etc.), ontological relationships between the features, demographics data and geographic codes, such as postal codes.

We have developed two GKB instances: the first has detailed information about the main Portuguese territory; the second, holds information about the main regions, countries, cities and places around the world in four different languages: Portuguese (PT), Spanish (ES), English (EN) and German (DE). While the first was created to support the GeoTumba service for Portugal, the latter is simply intended for validation of the GeoTumba software, through experiments with annotated multilingual corpora and Geo-IR evaluations covering other parts of the world, such as GeoCLEF.

GKB models geographic information as typed features and relationships. Figure 2 shows the feature types for geo-administrative and geo-physical domains data and their relationships as defined for the created world ontology. The main feature type is `ISO-3166-1`, which encompasses countries and territories. The feature types `Region`, `ISO-3166-2`, `Agglomeration`, `City-Capital`, `Place` and `Administrative Division` have a `'PartOf'` relationship with `ISO-3166-1`.

The relationship between `ISO-3166-1` and `Region` feature types is bidirectional, that is, a feature of type `ISO-3166-1` can be part of a `Region` (*Nicaragua* is part of *Latin America*) or a `Region` can be part of a `ISO-3166-1` feature (*Siberia* is part of *Russia*). The geo-administrative information is related to the geo-physical information through the feature types `ISO-3166-1` and `Region`. An instance of an `ISO-3166-1` can be part of `River`, `Continent` or `Lake`, or it may be adjacent to a `Sea`, `River` or `Lake`.

The geographic ontology of the world was built from two public information sources:

Wikipedia: on-line encyclopædia [21]. We used its name definitions of countries and their capitals in the four supported languages. We also collected all the geo-physical names information from this source. Names are defined in accordance to standards `ISO-3166-1` and `ISO-3166-2` (`http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html`).

World Gazetteer: information about the largest cities and agglomerations around the world [22]. We selected those with population above 100,000.

Table 1 shows some statistics for the world ontology used in GeoCLEF. The majority of the relationships are of the `PartOf` type, while `Equivalence` and `Adjacency` relationships are much less frequent. There are 12,283 features, and 7,970 of them have a population associated. Most of the features are provided by the World Gazetteer. Only relationship types `PartOf` and `Adjacency` are used to connect all features.

It is worth mentioning that the features `ISO-3166-1` contain preferred and alternative names, which includes the adjectives of the countries (e.g. Brazilian, Australian or Finnish). We loaded the preferred and alternative names in the four supported languages, while the adjectives were only loaded for the English and German languages.

For some types, the number of described features (number of Seas, Lakes and Regions) is much smaller than in reality because they weren't available in the information sources. We decided to add some features

Geographic Administrative Domain	
Number of features by types	Value
ISO-3166-1 (4 languages)	239
ISO-3166-2 (in English)	3,979
Agglomeration (in English)	751
Place (in English)	3,968
Administrative Division (in English)	3,111
City-Capital (4 languages)	233
Regions (4 languages)	2
Total number of features	**12,283**
features of population	7,970 (64,88%)
Features from Wikipedia	4,453 (36,25%)
Features from World Gazetteer	7,830 (63,75%)
Relationships PartOf	11,995
Relationships Equivalence	2,501
Total number of relationships	**14,496**
Geographic Physical Domain	
Number of features by types	Value
Planet (4 languages)	1
Continent (4 languages)	7
Sea (4 languages)	1
Lake (4 languages)	1
Total number of features	**10**
Features from Wikipedia	10 (100%)
Relationships PartOf	9
Total number of relationships	**9**
Inter-Domain Relationships	
Statistic	Value
Relationships PartOf	241
Relationships adjacency	13
Total number of relationships	**254**
Total	
Total number of features	**12,293**
Total number of relationships	**14,759**
PartOf relationships	12,245 (82,97%)
Equivalence relationships	2,501(16,95%)
Adjacency relationships	13 (0.08%)
Avg. broader features per feature	1.07
Avg. narrower features per feature	475.44
Avg. equivalent features per feature with equivalent	3.82
Avg. adjacent features per feature with adjacent	6.5
Features without ancestors	1(0.00%)
Features without descendants	12,045 (97,98%)
Features without equivalent	11,819 (96,14%)
Features without adjacent	12,291 (99,99%)

Table 1: Descriptive statistics of the world ontology

manually to GKB, because some of the GeoCLEF topics included place names like the North Sea, Caspian Sea and Siberia, which are not present on the information sources used to create the ontology.

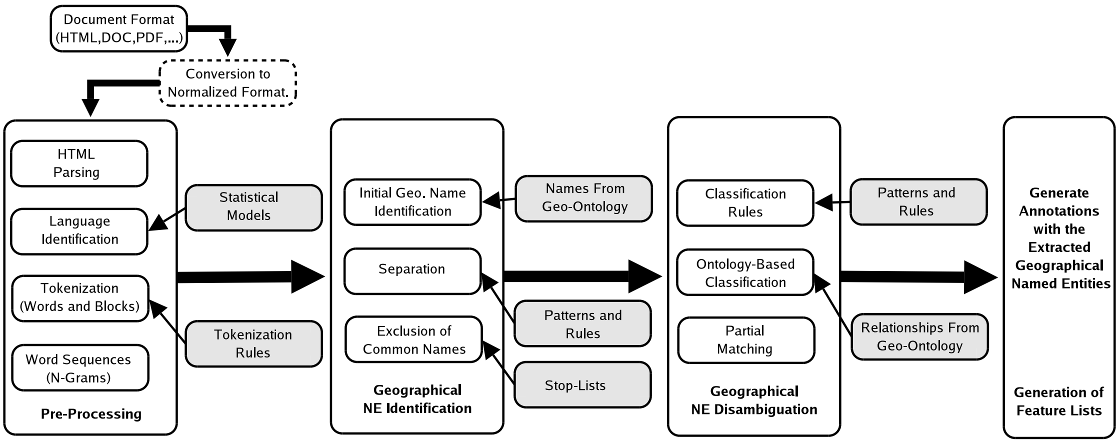
Document Format (HTML,DOC,PDF,...)
Conversion to Normalized Format.
HTML Parsing
Language Identification
Tokenization (Words and Blocks)
Word Sequences (N-Grams)
Pre-Processing
Statistical Models
Tokenization Rules
Initial Geo. Name Identification
Separation
Exclusion of Common Names
Geographical NE Identification
Names From Geo-Ontology
Patterns and Rules
Stop-Lists
Classification Rules
Ontology-Based Classification
Partial Matching
Geographical NE Disambiguation
Patterns and Rules
Relationships From Geo-Ontology
Generate Annotations with the Extracted Geographical Named Entities
Generation of Feature Lists

Figure 3: The geographical named-entity recognition and disambiguation step.

2.2 CaGE

CaGE is a text mining module specifically developed to infer the geographic context from collections of documents, based on the geographic knowledge presented in a OWL ontology generated by GKB. The main assumption of CaGE is that although each document can have several different geographic scopes (for instance, a news article covering several events in different places), we take the simplistic approach of associating each document to a single scope, or none if the assignment can not be made within a certain confidence level. This is similar to the "one sense per discourse" assumption, taken in many recognition and disambiguation systems [5]. Resources can nonetheless be characterized according to different degrees of locality, whether they are likely to be of interest to a geographically broad audience, or to relatively narrow regions. Our scopes can be seen as a sub-region hierarchy, where broad regions are detailed in their constituent locations.

The process of geo-referencing the textual documents is performed in two stages:

1. Identify the geographical references present in the document's text and weight them according to frequency.

2. Assign a corresponding geographical scope to each document according to the geographical references, their frequency, and the relationships among them.

The geographical references are handled through a named-entity recognition (NER) procedure particularly tailored to recognize and disambiguate geographical concepts over the text. Although NER is a familiar task in Information Extraction [17], this work advances the state of the art by presenting a specific adaptation strategy to the domain of multilingual geographical references. Besides recognizing place names, we try to normalize them in a way that specifically describes or even uniquely identifies the place in question, disambiguating them with respect to their specific type (e.g. city) and grounding them with features from the geographical ontology.

Figure 3 illustrates the geographical NER and disambiguation stage. It follows the traditional NER architecture by combining lexical resources with shallow processing operations. We have four processing steps:

Pre-processing: this step essentially performs text tokenisation. A language guesser is run on the document's text and this is the starting point for the following processing operations [13]. The identified textual segments are split into their constituent word *n*-grams, by moving a window over each text segment and taking all possible consecutive word sequences.

Named-entity identification: involves the detection of all possible *n*-grams that are likely to belong to a geographical reference. An initial identification applies language-specific patterns which combine

Expression Type	Expressions
Identifiers	city, municipality, district, street, avenue, river, island, mountain, valley, country, continent, zone, region, county, parish, desert, civil parish, hill, pueblo, village, villa, republic, peninsula
Contained	all over, among, amongst, at, in, inside, out of, through, throughout, within
Not Contained	around, outside, surrounding
Direction	above, across, along, ahead, backward, backwards, behind, below, beneath, beside, between, forward, forwards, in front, into, left, off, over, right, towards, towards, under
Relative Distance	adjacent, against, apart, away, beyond, close, closer, distant, far, farther, furthers, from, further, near, nearby, next, on
Earth Oriented	up, down, high, higher, low, lower, east, east of, north, north of, south, west, south east, south west, north east, north west
Complex Expressions	"cities such as", "districts like", "and other cities", "cities, including", "cities, especially", "one of the cities", and similar pattern for other place type identifiers

Table 2: Expressions used for recognizing geographical concepts in text.

place names from the geographical ontology and context expressions, with and without capitalization (i.e. "city of Lisbon" or "Lisbon metropolitan area"). Table 2 illustrates some of these expressions for English, and equivalents are used for the German language. Next, *n*-grams that are likely to contain more than one named-entity are detected and attachment problems are resolved. Finally, membership in exclusion lists is used to discard very frequent words that, despite having a geographical connotation, are more frequently used in other contexts (e.g. brand names or proper names).

Named-entity disambiguation: named-entity identification is not sufficient by itself to derive the meaning of expressions, as many named-entities remain ambiguous. This stage addresses this issue, aiming to find the correct meaning for the expressions recognized. Classification rules, built on the same expressions that are used to recognize entities, are first employed to solve the simple cases (e.g. in "city of X", we know X is a city and not some other geographical feature). Ontology based classification uses the feature types and other contiguity measures to guess the correct type for a given reference (i.e. a one referent per discourse assumption, so that place names throughout the same paragraph refer to the same or to geographically related locations). Finally, we compare slight word variations (i.e. one different character, one extra character or one less character) against references already disambiguated.

Generation of feature lists: this stage simply consists in listing the geographical references recognized in the text, together with their frequency and an association to the corresponding feature at the geographical ontology. In the cases not covered by the disambiguation heuristics, we use the associations to the several different possible concepts at the ontology, and some ambiguity problems can therefore persist at the end of this process.

In the scope assignment stage, besides the ambiguity problems which may still persist after the first stage, different (sometimes conflicting) geographical expressions may be associated with the same document. More than simply counting the most frequent references, we need to combine the available information and disambiguate further among the different possible scope assignments that can be made for each document. This is the idea behind the scope assignment approach, which relies on the existence of a graph where the particular relationships between geographical concepts are specified. The geographical ontology provides the needed information. We convert it to a graph representation, weighting different semantic relationships (edges) according to their importance (i.e., equivalence relationships are more important than hierarchical relationships, which in turn are more important than adjacency relationships) and weighting different geographical concepts (nodes) according to the feature weights computed at the previous step (see Figure 4).

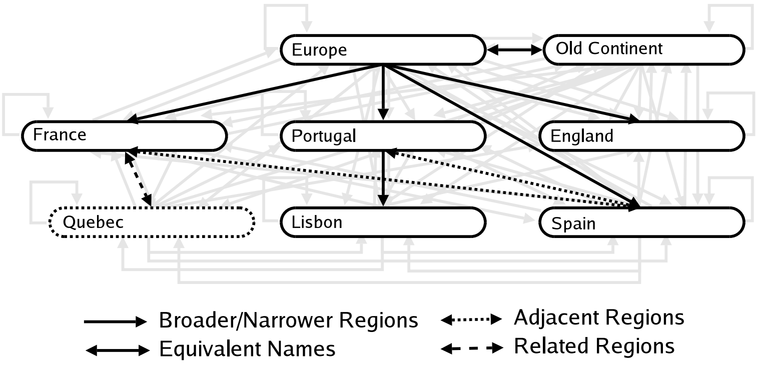

Figure 4: Geographic concepts graph

Importance scores are then calculated for all the nodes in the graph. In the end, we select the highest ranked node as the scope for the document. For the computation of importance scores, we use a variation of the popular PageRank ranking algorithm [16]. PageRank determines the importance of a vertex using the collective knowledge expressed in the entire graph, recursively computing importance using the idea of *voting*. The higher the number of *votes* (e.g. graph links) that are cast to a vertex, the higher its importance. Moreover, the importance of the vertex casting the vote determines the importance of the vote itself. There is a considerable amount of work focusing on all aspects of PageRank, namely stability, convergence speed, memory consumption, and the connectivity matrix properties [7]. By using this formulation we can leverage on all these previous studies.

PageRank is traditionally computed through iterative solution methods. Formally, let $G = (V,E)$ be a directed graph with the set of nodes V and the set of edges E, where E is a subset of $V * V$. For a given node V_i, let $In(V_i) \subset V$ be the set of nodes that point to it, and let $Out(V_i) \subset V$ be the set of nodes that V_i points to. The values w_{ij} correspond to weights given to the edges connecting nodes V_i and V_j, and s_i correspond to weights given to each node V_i (the source strengths). Below we show the formula for graph-based ranking that takes into account edge and node weights when computing the score associated with a node in the graph. The ranking score of a node V_i is defined as:

$$S(V_i) = (1-d)s_i + d * \sum_{V_j \varepsilon In(V_i)} \frac{w_{ij}}{\sum_{v_k \varepsilon Out(V_j)} w_{jk}} S(V_j)$$

The parameter d is a damping factor set to 0.85, integrating into the model the probability of jumping from a given node to another random node in the graph (e.g. having a document associated with a completely different feature than the ones we were able to extracted from it).

The source strengths s_i should be positive and satisfy the following condition:

$$|In(V_i)| = \sum_{j=1}^{|V|} s_i.$$

After a score is computed for each feature from the ontology, we select the most probable scope for the document, by taking the highest scoring feature, or none if all features are scored below a given threshold. The general procedure goes as follows:

1. Normalize the ranking scores obtained through the graph ranking algorithm.

2. If there are no features with a weight above the threshold, then no scope is selected.

3. From the set of features with the highest weight above the threshold:

   (a) If there is only one, return it as the scope for the document.

(b) If there is more than one feature, but one of them corresponds to a generally broader concept in the ontology, return this broader feature as the scope.

(c) If there is more than one feature, but they all have a common direct broader feature in the ontology, select this broader feature as the scope.

(d) If there is more than one feature and no common direct broader concept exists, use demographics data to select the scope corresponding to the highest populated geographical region.

2.3 Ranking documents with geo-scopes

In GeoTumba, we use geo-scopes to create new indexes supporting fast searches. The indexes integrate seamlessly with classic IR document weighting algorithms (TF × IDF, BM25, PageRank). For geographic ranking, a key concept is geo-scope similarity metrics [12]. Geographic similarity can be computed using the following criteria:

Relative position of geo-scopes on the concepts graph: two scopes may be equivalent, one may contain the other, they may overlap or may be adjacent to each other. We can say that document D_1 is more similar to query Q than D_2 if the area of overlap between the geo-scopes of D_1 and Q is larger than the area of overlap between D_2 and Q.

Geometric proximity of geo-scopes: assuming that the similarity between two geo-scopes is inversely proportional to the distance between them, we can use Euclidean distance, travel time on the public transportation network or other source of data related to physical proximity to compute similarity as the distance between the centroids of two scopes.

Graph-based distances: the relative location of nodes representing scopes on the geographic concepts graph can be used to define a semantic similarity metric between scopes [9].

Importance of scopes: the total population or economic importance of the geographic entity represented by a scope can be used as criteria for weighting the relative power of a scope.

As we put emphasis on geographic reasoning rather than on geometric reasoning, the indexes used for geographic ranking on GeoTumba only have information about the similarity values between any pair of scopes, with little or no spatial information.

The best strategies for efficiently organising this information for fast access are still in debate. We are presently pondering strategies for fusing various similarity metrics in a global GeoTumba ranking function. As a result, we decided to participate in GeoCLEF with a system software configuration that does not use the geographic indexes, but still ranks documents according to geographic criteria.

2.4 Software Configuration used to create the GeoCLEF Runs

In our GeoCLEF experiments, we used QuerCol, a query expansion component to generate queries from the CLEF supplied topics (more details about the query generation process are presented in a separate text describing our participation in the CLEF 2005 ad hoc task [1]). QuerCol provides multiple strategies to generate queries from topic data to be submitted to search engines. For GeoCLEF, we used in some runs the location part of the topics as additional query terms. For bilingual runs, we asked our colleagues from the Braga node of Linguateca to translate the topics, using the same method as in our submission to the ad hoc bilingual subtasks.

For generating the runs submitted to GeoCLEF, we assembled along with QuerCol a modified retrieval system from our existing software. The changes made to GeoTumba include:

- Replacement of the web crawler by a custom loader already used in previous evaluations, to bring the GeoCLEF newswire text collections into the repository. We also used the language information provided with the document collection metadata and turned off the language guesser of CaGE.

- Development of a simple alternative scope assignment algorithm, that consists in simply selecting the most frequent geographical reference as the scope of a document. We were interested in comparing the graph-ranking method for assigning geographical scopes to documents against this baseline approach.
- Implementation of a geo-ranking function which does not use geographic indexes. Ranking was computed in two stages: first, we ranked documents using the classic IR approach, with a simplified version of the BM25 function that only uses the term index (this function was also used in the CLEF 2005 ad hoc task participation). Then, we ranked the documents in the obtained result set with a geographic similarity function. The final ranking corresponds to the set of documents containing the terms on the query, ordered by a composite key having the geographic rank followed by the non-geographic rank.

The geographic similarity metric that we used in GeoCLEF is defined on a *scopes tree* extracted from the geographic concepts graph built from the geographic ontology. The scopes tree contains the nodes of the graph and the edges defining `partOf` relationships among the nodes. In this tree, we define i) $depth(X)$ as the count of edges between node X and the root of the tree; ii) $ancestor(X,Y) = true$ if X is on the path of Y to the root node of the tree; and iii) TD, tree depth, the maximum $depth()$ of any node on the tree.

Given a query Q, a geo-scope $Scope_Q$ and a result set with documents $D_1,...,D_n$, each with a $Scope_{D_i}$ or `NULL` scope assigned, the geographic similarity $GS(Q,D_i)$ is obtained as follows:

$$GS(Q,D_i) = \begin{cases} 0 & if Scope_Q = Scope_{D_i} \\ depth(Scope_Q) - depth(Scope_{D_i}) & if\ ancestor(Scope_Q, Scope_{D_i}) = true \\ n \times TD + depth(Scope_{D_i}) - depth(Scope_Q) & if\ ancestor(Scope_{D_i}, Scope_Q) = true \\ 2 \times n \times TD & otherwise \end{cases}$$

The definition above means that the geographic similarity ranking function first ranks all the documents with the same scope as the query, then those with a narrower scope than the query, and then those with a wider scope. Finally, documents with `NULL` scopes or scopes that can not be defined as strictly narrow or wider than the scope of the query are ranked last.

3 Runs Description and Evaluation Goals

Our goals for participating in GeoCLEF were:

Scope ranking: measure how the ranking with the geo-scopes assigned to documents improves Geo-IR results, in comparison to include location terms in the query strings, using geographic terms as common terms, a common practice for narrowing geographic searches (e.g. '*restaurant london*') [8, 14].

Scope assigning: when using geo-scopes, compare the graph-based algorithm against the simple scope assignment algorithm that selects the most frequent geographic entity in texts.

Expansion of location terms: when not using geo-scopes, measure the contribution of the expansion of geographic terms in queries to improve searches.

Topic translation: observe the performance of Portuguese to English bilingual runs. Our efforts were focused towards the English monolingual subtask. The bilingual runs obtained provide initial results on the performance of the machine translation system being developed by the Linguateca group at Braga, Portugal. There was no interest in creating runs derived from manual queries for this subtask.

We submitted six runs for the English monolingual subtask, four runs for the German monolingual subtask, and four runs for the Portuguese to English bilingual subtask [20]. Table 3 summarizes the submitted runs. Below, we describe the creation procedures and observations intended for each of the submitted runs:

Run description	Monolingual EN	Monolingual DE	Bilingual PT->EN
(Mandatory) Automatic query generation, title + description only	XLDBENAutMandTD	-	XLDBPTAutMandTD
(Mandatory) Automatic query generation, title + description + location	XLDBENAutMandTDL	-	XLDBPTAutMandTDL
Manual query generation, title + description only	XLDBENManTD	XLDBDEManTD	-
Manual query generation, title + description + location	XLDBENManTDL	XLDBDEManTDL	-
manual query, title + description run, GKB 'PageRank'-like scopes	XLDBENManTDGKBm3	XLDBDEManTDGKBm3	XLDBPTManTDGKBm3
manual query, title + description run, most frequent NE scopes	XLDBENManTDGKBm4	XLDBDEManTDGKBm4	XLDBPTManTDGKBm4

Table 3: The runs submitted by the XLDB group to the GeoCLEF

'AutMandTD and AutMandTDL': GeoCLEF required two fully automatic mandatory runs. The first should use only title and description information from the supplied topics, while the second should use the location information as well. These two runs provide the evaluation baselines. The first indicates the performance of the non-geographical IR mechanisms being used, and the other provides the means to evaluate geographical IR against a simple baseline. The generated query strings contained the title terms, a maximum of 3 terms from the description, and the location terms in the case of '*AutMandTDL*'.

'ManTD': this run was generated as an intermediary step for the construction of the *ManTDL*, *TDGKBm*3 and *TDGKBm*4 runs. It provides a comparative baseline for the other submissions. We created manual queries to generate these runs, using terms from the topics's titles and descriptions, avoiding narrative terms and all related geographic terms. We did not include any location names or adjectives from the topics titles in the queries. We expanded morphologically the terms, and combined them using 'AND' and 'OR' logic operators into a single query line. As our baseline runs, the goal was to maximize recall. Precision was expected to suffer due to the lack of geographic terms on these baseline runs. These runs have a label which ends with '*ManTD*' (MANual query, Title + Description).

'ManTDL': this run was meant to measure the efficiency of a simple geographical IR technique, which consists in restricting the search to documents containing the geographical location terms provided in the topics, expanding also the geographical names in order to account for equivalent names and spatial operators. We wanted to measure how efficient this technique is for restraining document scopes, hence we created these runs by inserting the scope(s) location(s) from the topic to the manual query from the '*ManTD*' runs.

When the topic location scope implicitly embraces a group of countries, we extended it to the country level. For example, in the topic with the North Sea scope, the generated query string included terms like *North, Sea, England* and *Denmark*. In the case of topics with important spatial relation (e.g. South-West of Scotland), we expanded the scope in a similar way for each location found on the narrative, like *Ayr* and *Glasgow* on the example above (notice that this was the only information used from the narratives, regarding all query strings). These runs have a label which ends with '*ManTDL*' (MANual query, Title + Description + Location).

'TDGKBm3 and TDGKBm4': in this run, we intended to measure the efficiency of our text mining software for assigning documents with a corresponding geographical scope, as described in Section 2. Runs labeled with '*TDGKBm*3' mark the PageRank-like scope assignment, and the labels '*TDGKBm*4' mark the most frequent geographic entity as the scope's document.

We did not submit mandatory runs for the German monolingual task, because QuerCol couldn't handle the agglutinated concepts in the topic titles properly. We found no interest in submitting these runs as the German language specificities were outside the scope of our participation in GeoCLEF.

4 Results

the obtained results are presented in average precision vs. interpolated recall charts, in Figures 5 and 6 (English and German monolingual subtasks), and in Figure 7 (Portuguese to English bilingual subtask).

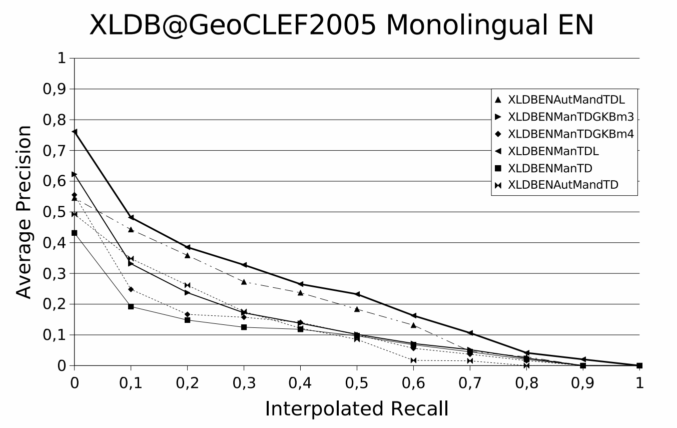

Figure 5: Results of the XLDB group on the English monolingual subtask of GeoCLEF 2005

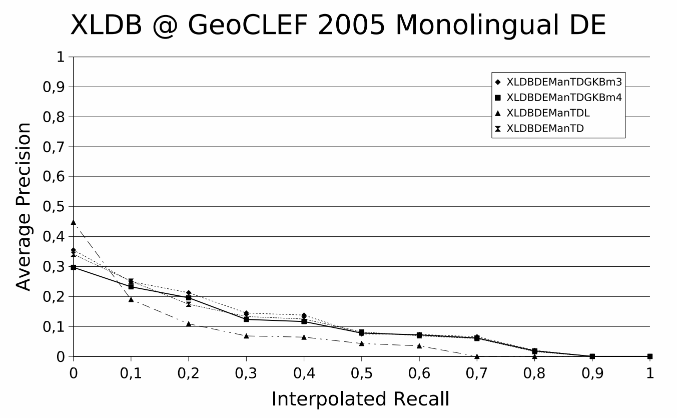

Figure 6: Results of the XLDB group on the German monolingual subtask of GeoCLEF 2005

Regarding the evaluation goals presented on the previous Section, we can derive from the observation of Figures 5, 6 and 7 the following conclusions:

Scope ranking: comparing no-scope runs vs. scope-aware runs, we observe that the runs with location terms inserted in the fully automatic query (*AutMandTDL*) ended with better precision than the runs with geographic scope ranking (*TDGKBm3* and *TDGKBm4*). We didn't expect this behaviour, as our Geo-IR is able to retrieve relevant documents to a given scope without its name on the query. A more detailed analysis of the qrels shows that this happened because both the geo-ranking method and the ontology data revealed some limitations.

Scope assigning: comparing the graph-based vs. the most frequent geographical reference algorithms used to assign scopes to documents, the method based on the graph ranking algorithm (*TDGKBm3*) achieved higher precision than the alternative method of assigning the most frequent geographic reference as the document's scope (like the *TDGKBm4* runs). Scrutinizing the results, we can see

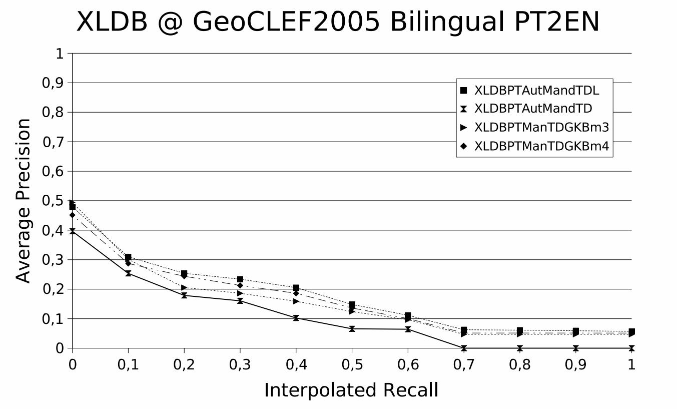

Figure 7: Results of the XLDB group on the Portuguese to English bilingual subtask of GeoCLEF 2005

that CaGE normally assigned the same scopes that an human would infer if he only had the same geographic knowledge passed on the world ontology.

Expansion of location terms: We can observe that the runs based on manual queries with expanded location terms (i.e. the *ManTDL* runs) obtained higher precision than the *AutMandTDL* runs. This reinforces our belief that relevant documents often do not contain explicitly the terms from the desired location. A Geo-IR system should consider the relationships between geographical concepts in order to retrieve relevant documents to a given location, even if they do not contain the location terms. However, the CaGE graph-ranking algorithm did not obtained better results than the runs created by using only location names and a standard text search (*AutMandTDL*). As scopes seemed to be correctly assigned, we suspect the result was due to the used ontology and geographic ranking function.

Topic translation: The English monolingual runs exhibit better results than the bilingual runs. This was due to the quality of the topic translation, which faced some difficulties. Detailed description of these problems are included in the ad hoc participation paper [1]. This wasn't too obvious on the *ManTD* runs (they showed a similar performance), as they were created from query strings with few terms selected from the topic.

The analysis of the topic qrels shows that 61% of the relevant documents have been assigned to an unrelated or unknown scope. We realized that sub-optimal results are caused by the geographic ranking strategy adopted, and the lack of relationships in the ontology. For example, we have '*Glasgow*' as part of '*United Kingdom*', and '*United Kingdom*' as part of '*Europe*'. Yet, the record '*Scotland*' was associated to '*United Kingdom*', and thus our geo-ranking module did not have a path from '*Glasgow*' and '*Scotland*' on the scopes tree.

Further analysis also revealed that we could have profited from using the `Adjacency` relationships on the geographic similarity metric, as we couldn't associate documents with assigned scopes like *Russia* or *Azerbaijan* to regions like *Siberia* or *Caspian Sea*.

These facts had a noticeable impact on the *TDGKBm3* and *TDGKBm4* runs, meaning that we can't make an overall evaluation of our Geo-IR, compared to the *AutMandTDL* and *ManTDL* runs, at this point.

5 Conclusion

For our participation at the GeoCLEF evaluation campaign, we adapted software from a geographical web search engine currently under development at our group. Our approach is based on a two stage process, in which geographical references in the text are recognized and a geographic scope is afterwards computed for each document. A central component of the whole process is a geographical ontology, acting as the source of geographical names and relationships.

Although our scope assignment algorithm has shown to be better than a simple baseline of selecting the scopes according to the most frequent geographical references, retrieving documents using scopes was no better than the simple inclusion of the topic locations as additional terms to a standard text search. Our evaluation of the qrels has shown that the lack of information about some of the geographic concepts and their relationship to other concepts on the ontology that we built was the cause for very poor performance in a considerable number of topics. This shows that the success of our approach strongly depends on the amount and quality of geographic knowledge that is provided to the system. However, we suspect that if too much detailed geographic information is provided, performance will also become sub-optimal.

A similar resource to GKB is the Getty Thesaurus of Geographic Names (TGN) [15], which is a structured vocabulary including names and associated information about both current and historical places around the globe. The focus of TGN records are places, each identified by a unique numeric ID. Linked to the place's records are names (historical names, common alternative names and names in different languages), place types (e.g., inhabited place and state capital), place's parent or position in the hierarchy, other relationships, geographic coordinates, notes and the data sources. There may be multiple broader contexts, making the TGN poly-hierarchical. In addition to the hierarchical relationships, the TGN has equivalent and associative relationships, similar to the GKB structure. We believe that the number of features in GKB is enough to assign the geographic scope to each document. We wanted to experiment this assumption with other gazetteers, and we planned to generate runs using TGN to compare the results to the ones obtained with GKB, but we did not receive it in time to be used at GeoCLEF.

As future work, in addition to improving the scope assignment algorithm and experimenting with more comprehensive ontologies, we plan to devise and evaluate better geographic ranking functions, capable of geographically ranking documents even in the absence of geographic knowledge about terms of the query location part or in documents, and making better use of the geographic scopes.

6 Acknowledgements

We thank Andreas Wichert, for the manual creation of the German queries and insight on certain aspects on diacritics expansions on German texts, and Alberto Simões, for the translation of the topics.

Thanks to all the tumba! developers and GREASE project participants. This work was financed by the Portuguese Fundação para a Ciência e Tecnologia through grant POSI / PLP / 43931 / 2001 (Linguateca) and by grant POSI / SRI / 40193 / 2001 (GREASE). Bruno Martins is supported by FCT through grant SFRH-BD-10757-2002.

References

[1] Nuno Cardoso, Leonardo Andrade, Alberto Simões, and Mário J. Silva. The XLDB Group participation at CLEF 2005 ad hoc task. In C. Peters, editor, *Working Notes for the CLEF 2005 Workshop*, Wien, Austria, 21-23 September 2005.

[2] Nuno Cardoso, Mário J. Silva, and Miguel Costa. The XLDB Group at CLEF 2004. In C. Peters, editor, *Working Notes for the CLEF 2004 Workshop*, Bath, UK, 15-17 September 2004.

[3] Marcirio Chaves, Mário J. Silva, and Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. In *20^o Simpósio Brasileiro de Banco de Dados - SBBD, Uberlândia*, 3rd - 7th October 2005.

[4] Marcirio Silveira Chaves, Mário J. Silva, and Bruno Martins. GKB - Geographic Knowledge Base. Technical Report DI/FCUL TR 5-12, June 2005.

[5] W. Gale, K. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, 1992.

[6] GREASE home page. http://xldb.di.fc.ul.pt/index.php?page=GREASE.

[7] Ashraf Khalil and Yong Liu. Experiments with PageRank computation. Technical Report 603, Computer Science department at Indiana University, December 2004.

[8] Janet W. Kohler. Analysing search engine queries for the use of geographic terms. Master's thesis, University of Sheffield, September 2003.

[9] Yuhua Li, Zuhair Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, 2003.

[10] Bruno Martins and Mário J. Silva. Geographical named entity recognition and disambiguation in web pages. 2005. (to appear).

[11] Bruno Martins and Mário J. Silva. A graph-based ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining*, 2005.

[12] Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in Geo-IR systems. In *Workshop on Geographical Information Retrieval, CIKM 2005*, November 4th 2005.

[13] Bruno Martins and Mário Silva. Language identification in Web pages. In *Proceedings of ACM-SAC-DE-05, the Document Engineering Track of the 20th ACM Symposium on Applied Computing*, 2005.

[14] Bruno Martins and Mário J. Silva. A Statistical Study of the WPT 03 Corpus. Technical Report DI/FCUL TR-04-1, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, April 2004.

[15] Getty Thesaurus of Geographic Names (TGN). http://www.getty.edu/research/conducting_research/vocabularies/tgn/.

[16] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library, November 1999. Working Paper.

[17] Tjong Kim Sang, Erik F., and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003, the 7th Conference on Natural Language Learning*, pages 142–147. Edmonton, Canada, 2003.

[18] Mário J. Silva. The Case for a Portuguese Web Search Engine. In *Proceedings of ICWI-03, the 2003 IADIS International Conference on WWW/Internet*, pages 411–418, Algarve, Portugal, 5-8 November 2003. IADIS.

[19] Mário J. Silva, Bruno Martins, Marcirio Chaves, Nuno Cardoso, and Ana Paula Afonso. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems*, (accepted for publication).

[20] GeoCLEF task description. http://ir.shef.ac.uk/geoclef2005/task_description.html.

[21] Wikipedia. http://www.wikipedia.org.

[22] World Gazeteer. http://www.world-gazetteeer.com.