B. Martins*, H. Manguinhas*, J. Borbinha*, W. Siabato**

A geo-temporal information extraction service for processing descriptive metadata in digital libraries

Keywords: Georeferencing; gazetteers; geoparser; digital libraries; DIGMAP.

Summary

In the context of digital map libraries, resources are usually described according to metadata records that define the relevant subject, location, time-span, format and keywords. On what concerns locations and time-spans, metadata records are often incomplete or they provide information in a way that is not machine-understandable (e.g. textual descriptions). This paper presents techniques for extracting geotemporal information from text, using relatively simple text mining methods that leverage on a Web gazetteer service. The idea is to go from human-made geotemporal referencing (i.e. using place and period names in textual expressions) into geo-spatial coordinates and time-spans. A prototype system, implementing the proposed methods, is described in detail. Experimental results demonstrate the efficiency and accuracy of the proposed approaches.

Introduction

Previous studies showed that geographic and temporal criteria both have important roles in filtering, grouping and prioritizing information resources [2][26][21], motivating research in methods for transforming human-made geo-temporal references (i.e. place or time-denoting expressions) into machine-understandable representations (i.e. geo-spatial coordinates and intervals in a calendar system). Geo-temporal information extraction concerns the automated process of a) analyzing text, b) finding and disambiguating geographic and temporal references, and c) combining these references into meaningful semantic summaries (i.e. geotemporal scopes for the documents). The text may come from Web pages, from resources in content management systems, or from metadata records in digital libraries. The problem has been addressed with mixed success, for instance by the Natural Language Processing [3] and Geographical Information Retrieval [22][2] communities.

This paper describes automated methods for extracting geo-temporal information from text, using relatively simple text mining methods that leverage on a Web gazetteer service [6].

The proposed techniques are evaluated through comparisons with a gold-standard collection of textual resources, where each item has a geo-temporal context assigned by humans. The evaluation collection consists of metadata records from the DIGMAP[1] digital library of old maps [7], having temporal and geographical annotations provided by librarians.

The paper also presents a prototype geo-parser system[2], developed in the context of DIGMAP and demonstrating the proposed techniques. The geo-parser can process plain-text or XML documents, extract the geo-temporal information, and output the results in XML. Through this geo-parser, the metadata records can be augmented with machineunderstandable geo-temporal information, leveraging on XML

* Instituto Superior Técnico - Department of Computer Science and Engineering. Av. Rovisco Pais, 1049-001 Lisboa, Portugal [bruno.g.martins@ist.utl.pt] [hugo.manguinhas@ist.utl.pt] [jlb@ist.utl.pt]

** Universidad Politécnica de Madrid – Laboratory of geographic information technologies (LatinGEO) Campus Sur UPM. Km. 7.5 Autovía de Valencia, 28031, Madrid, España [wsiabato@acm.org]

[1] www.digmap.eu

[2] http://www.digmap2.ist.utl.pt:8080/geoparser/

time and location extensions that already widely deployed, e.g. OGC's Geography Mark-up Language (GML)[3].

The rest of this paper is organized as follows: Section 2 presents the main concepts and related works; Section 3 presents the proposed techniques for geo-temporal information extraction; Section 4 describes the geo-parser Web service developed in the context of DIGMAP, also describing a prototype interface for exploring geo-temporal information over maps and timelines; Section 5 presents results from evaluation experiments; finally, Section 7 presents our conclusions and directions for future work.

Concepts and related works

Extracting different types of entities from text is usually referred to as Named Entity Recognition (NER). For at least a decade, this has been an important natural language processing task [9]. NER has been successfully automated with near-human performance. However, the work described here differs from the standard NER task:

- The types for our named entities (e.g. references to cities or villages) are more accurate than the course-grained types that are generally considered (i.e. person, organization or location).
- The documents are multilingual and we may have to address languages for which annotated corpora are scarce (e.g. Portuguese or Spanish). As in other text mining tasks, more NER work has been done for English.
- Recognition in itself does not derive a meaning for the recognized entities, and we must also match them explicitly to spatial areas and time-spans (i.e. match the references to exact gazetteer entries). Extending NER with gazetteer matching presents harder problems than the simple recognition [17].
- Handling large collections requires processing the individual resources in a reasonable time, constraining the choice of techniques and heuristics. Performance issues were often neglected in previous NER evaluation studies.
- The named entities in a text can be seen as part of a specific semantic context. These entities should be combined into meaningful semantic summaries (i.e. an encompassing geo-temporal scope for each document), taking into account the relationships among them (e.g. part-of relationships).

Traditional NER systems combine lexical resources (i.e. gazetteers) with shallow processing operations, consisting of at least a tokenizer, a lexicon and NE extraction rules. Tokenization segments text into tokens, e.g. words and punctuation. The rules for NE recognition are the core of the system, combining names in the lexicon with elements like capitalization and surrounding text. These rules can be generated by hand or automatically, through machine learning. The former method relies on experts, while the latter induces rules from manually annotated training data.

The best machine learning systems achieve f-scores over 90% in newswire texts. However, they require balanced and representative training corpora [20]. A bottleneck occurs when such data is not easily available. This is usually the case with non-English languages or very specific tasks, such as recognizing and disambiguating thin-grained geo-temporal references. The degree to which gazetteers help in identifying named entities also seems to vary. While some studies showed that gazetteers did not improve performance [16], others reported significant improvements using gazetteers and trigger phrases [11]. Mikheev et al. showed that a NER system without a lexicon could perform well for most classes, although not for places [19]. The same study also showed that simple gazetteer matching performs reasonably well. Eleven out of the sixteen teams at the NER shared task of the 2003 Conference on Com-

[3] http://www.opengis.net/gml/

putational Natural Language Learning (CoNLL-2003) used gazetteers in their systems, all obtaining performance improvements [20].

An important conclusion of CoNLL-2003 was that ambiguity in geographic references is bidirectional. The same name can be used for more than one location (referent ambiguity), and the same location can have more than one name (reference ambiguity). The same name can also be used for locations and other entity classes, such as persons or company names (referent class ambiguity). A recent study estimates that more that 67 percent of the place references in a text are ambiguous [13]. Another study shows that the percentage of place names that are used by more than one place ranges from 16.6 percent for Europe to 57.1 percent for North and Central America [28].

A past workshop addressed techniques for exploring place references in text, focusing on more complex tasks than the simple recognition [3]. Some of the presented systems addressed the full disambiguation of place references (i.e. geo-parsing) although only initial experiments have been reported. The usual architecture for these systems is an extension of the general NER pipeline, adding stages that address the matching of the extracted nams to gazetteer entries (see Figure 1).

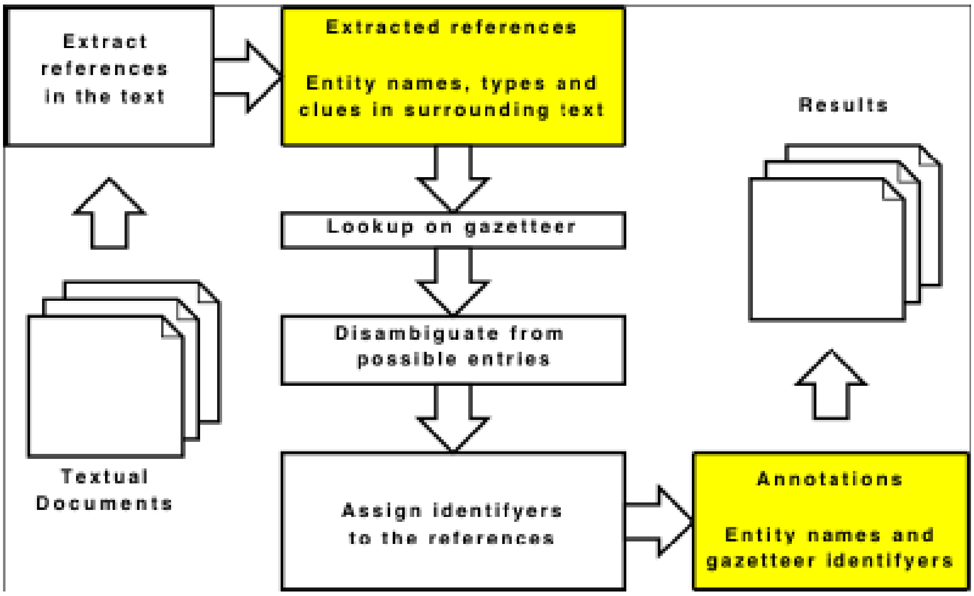

Extract references in the text
Extracted references
Entity names, types and clues in surrounding text
Results
Lookup on gazetteer
Disambiguate from possible entries
Textual Documents
Assign identifyers to the references
Annotations
Entity names and gazetteer identifyers

Figure 1. Typical approach for geo-parsing text

In order to find the correct sense of a geographic reference systems usually use plausible heuristics [15]:

- **One referent per discourse:** an ambiguous geographic reference is likely to mean only one of its senses when used multiple times within one discourse context (e.g. the same document). This is similar to the one sense per discourse heuristic proposed for word sense disambiguation [12].
- **Related referents per discourse**: geographic references appearing in the same discourse context tend to indicate nearby locations. This is an extension of the heuristic presented in the first point.
- **Default senses**: a default sense can be assigned to ambiguous references, as important places are more likely to be referenced (e.g. the name Lisbon is more likely to reference a city than a street).

Research into geo-parsing approaches is only now getting momentum. A good survey was given in [23] but, in comparison with standard NER, considerably less information is available. Different combinations of the three heuristics above have been tested [13][23], but results are difficult to compare. The systems vary in the types of classification and disambiguation performed, and the evaluation resources are also not consistent [10][23].

Regarding interoperability, the Open Geospatial Consortium (OGC[4]) already proposed a simple Web Geo-parsing Service for recognizing place references. However, this document is currently discontinued [4]. Although providing comprehensive details on the service interface, the document did not discuss any issues related to implementation. SpatialML[5] is another recent proposal for interoperability between geo-parsing systems, emphasizing the need for standard evaluation resources. The prototype system reported in this paper uses an XML format similar to the one proposed by the OGC, with extensions related to the temporal references and to the association of place references to geo-spatial coordinates.

Previous works have also addressed the combination of place references given in a text in order to find the encompassing geographic scope that the document discusses as a whole. For instance Web-a-Where proposes to discover the geographic focus of Web pages using part of relations described in a gazetteer [5] (i.e. Lisbon is part of Portugal, and documents referencing both these places should probably have Portugal as the scope). Looping over a set of disambiguated place references, Web-a-Where aggregates for each page the importance of the various levels a gazetteer hierarchy. These taxonomic levels are then sorted by score and results above a given threshold are returned as the page focus. In Web pages from the ODP[6], Web-a-Where guessed the correct continent, country, city, and exact scope respectively 96, 93, 32, and 38 percent of the times. More advanced methods have also been described [19], but at the cost of additional complexity and computational efforts.

On what concerns temporal references, previous reports have addressed the linking of events with time and the ordering of events [8][13]. Similarly to the case of places, there exists a precise system for specifying time and time ranges (i.e. calendar systems), but people often use ambiguous names instead [24]. Ambiguity in temporal references is perhaps even a bigger challenge that in the case of places, particularly for applications requiring fine-grained temporal annotations (e.g. Easter comes in different dates for the Catholic and Orthodox churches, Winter depends on the hemisphere, etc.). The work reported in [8] described an approach for deep time analysis, capable of satisfying the needs of advanced reasoning engines. The approach was rooted on TimeML, an emerging standard for the temporal annotation of text that defines an XML format for capturing properties and relations among time-denoting expressions [14]. However, in our work, we are only addressing temporal references at a much simpler level. We never try to disambiguate expressions such as Monday or Autumn, instead focusing on complete dates and on names for historical periods. At most, we deal with reference/referent ambiguity issues similar to the ones that are present in the case of place names (e.g. the term Renaissance can indicate a period in the twelfth century in Europe or other periods), again by using heuristics.

An important work addressing the combined analysis of temporal and geographic references is ECAI TimeMap. This project addressed the exploration of scholarly materials in space and time [24]. Particular attention was given to the development of geo-temporal gazetteers [25], but issues related to information extraction were not the main focus of the project.

Geo-temporal text mining

This section presents the proposed techniques for geographical and temporal information extraction. We start by presenting the gazetteer service, followed by the algorithms for geographical and temporal information extraction. Finally, we present general issues related a Web geo-parser service implementing the proposed approaches.

[4] http://www.opengeospatial.org

[5] http://sourceforge.net/projects/spatialml

[6] http://www.dmoz.org/

The geo-temporal gazetteer

Having a multilingual gazetteer with comprehensive information about names of places and historical periods, together with their properties (i.e. place types, spatial coordinates, time spans, hierarchical position, alternative names and semantic associations) is a key requirement to our task. Figure 2 illustrates the most important gazetteer elements.

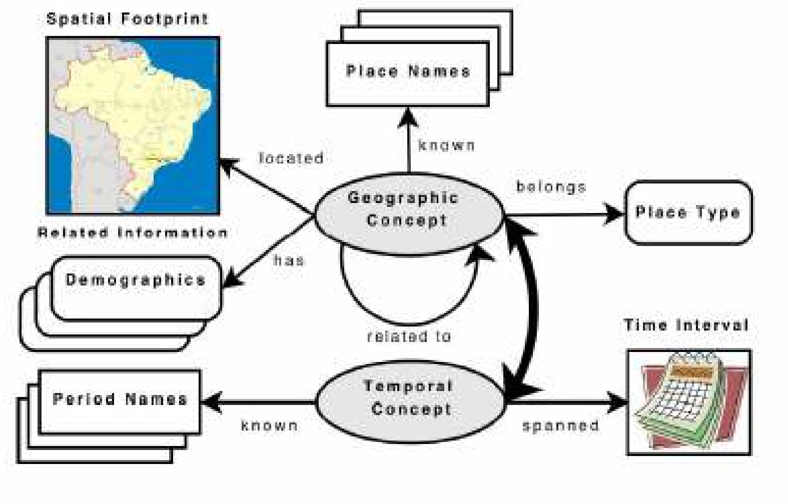


Figure 2. Core elements of the DIGMAP gazetteer

In the context of DIGMAP, we developed a Web gazetteer service integrating data from multiple sources (e.g. the GeoNames[7] website or the ECAI time period directory [25]). This gazetteer follows the XML Web interface and the data model proposed for the Alexandria Digital Library (ADL) gazetteer [1], introducing changes related to the handling of temporal data. A separate publication describes the gazetteer in more detail [6] and Table 1 provides basic descriptive information. The data in the gazetteer directly influences the outcome our experiments, as we depend on it for the correct interpretation of the text references. In particular, the disambiguation methods depend on a rich set of hierarchical relations among the gazetteer entries.

Statistic	Value	Comment
Number of places	7.034.538	Mostly modern geography
Number of place names	15.026.983	
Number of place types	210	Preferred terms ADL-FTT
Number of historical periods	1989	ECAI Time Period Directory + Wikipedia
Places with spatial footprints	6.621.138	Mostly centroids, and a few bounding boxes
Number of relationship types	5	
Number of places with relations	431.397	Still missing the relations from GeoNames relations
Number of place	866.019	Mostly part-of and contains
Number of time/place relations	1989	

Table 1. Statistical characterization of the DIGMAP gazetteer.

[7] www.geonames.org

We are currently extending the gazetteer by integrating data from other sources of place data besides GeoNames namely authority records from library catalogues and other gazetteer sources. A complete description of the gazetteers we are evaluating to include in the Web gazetteer service were published in the third DIGMAP newsletter[8].

Extracting geographical information

The extraction of geographical information is divided in three parts, namely the recognition of place references, the disambiguation of place references and the combination of combination of place references into geographic scopes.

Recognition of place references

For the recognition of place references we use a simple NER method based on word tokenization and gazetteer lookups. Capitalized names, i.e. sequences of n consecutive words with the first letter in upper case, are first matched against a list of important places, i.e. places listed in the gazetteer with a type class above a given threshold. For each class in the gazetteer's hierarchical classification system we assign a score $s \in [0,1]$ (e.g. continents are more important that countries, countries are more important that cities, cities are more important than villages, etc.). Place names with a type class having a score $s \leq 0.5$ are discarded from the list used for the simple recognition (i.e. places bellow city). The rationale is that names for small unimportant places are highly ambiguous, but very simple techniques can be effective for recognizing the names of large and important geographical areas.

For names of small regions we use look-ups in a separate list, containing all place names in the gazetteer. However, instead of the simple matching procedure, we also look for the presence of words indicating place types (e.g. words like district or municipality) in the surrounding text. These surroundings are given by a window of 3 words appearing before and after the recognized place reference. We only consider the place reference if the name is accompanied by a place type. The surrounding text approach is general enough to work for several different languages, only requiring a multilingual list for words that correlated with specific place types.

Disambiguation of place references

For the disambiguation of place references we use queries to the gazetteer together with a simple set of heuristics. From the previous step, the names for small geographic regions already contain an associated place type. For the other names, we also look in the surrounding text for words indicating place types, although the recognition does not depend on finding them. In the cases in which we have place type information, the query to the gazetteer combines the name with the type, and only features having that exact combination are returned. In the case of names having no associated place type, we simply query the gazetteer for features having the same exact name.

After the gazetteer returns geographic features matching the query, we rank them according to a score reflecting a default sense heuristic. For a given feature f, and in the case of names having an associated place type, the score is given by the normalized count of the number of child features for f that are defined in the gazetteer. The idea is that places with more subdivisions are more likely to be referenced. In the case of a name without an associated feature type, the score corresponds to the previously defined s value associated with the place type of the gazetteer feature f, on the rationale that features with a higher s score are more likely to be referenced.

Finally, for the place references having more than one matching feature returned by the gazetteer query, we use the one reference per discourse and related referents per discourse assumptions to try adjusting the ranking scores. The features having a parent or child feature found in the set of all references discovered in the document have their ranking score boosted by 0.2 up to a maximum value of 1.

[8] http://www.digmap.eu/doku.php?id=wiki:digmap_newsletter

The end result of this stage is a set of place references and, for each, a list of possible referencing concepts ordered by their ranking score. This score can be seen as the probability of a given place reference being indeed related to that particular feature.

Combining the disambiguated place references

After recognizing and disambiguating place references we combine them in order to find the general geographical context covered by the document. This is done through a similar technique to the one proposed in [5].

For each feature potentially referenced in the document, we use the part-of relationships defined in the gazetteer to fetch the hierarchical parents up to the root level. This results on a set S of possible geographic scopes.

In this set S, all the features that are referenced in the document start with the ranking score given in the extraction stage. These scores are then propagated and aggregated into the hierarchical ancestors, using a quadratic function to decrease the propagated ranking score according to the hierarchic level. For instance, let us consider a document containing references to geographical features *A* and *B*, with corresponding scores s_A and s_B. Let us also consider that the gazetteer contains part-of relationships corresponding to the hierarchies *C/B/A* and *C/B*. The final score of feature A would be s_A, for feature *B* would be $s_B+s_A*0.75$ and for feature C would be $s_B*0.75+s_A*0.752$. The end result of this stage is a list of possible geographic scopes (i.e. the features in S with a score grater than zero) ordered by the corresponding aggregated score. This score can be seen as the probability of a given geographic feature representing the geographic scope of the document.

Extracting temporal information

Similarly to the geographical case, the extraction of temporal information is also divided in three parts, namely recognition, disambiguation and scope assignment.

The recognition stage again uses a simple NER method based on word tokenization and gazetteer lookups. Capitalized names are matched against a list containing names of historical periods. This approach is complemented with regular expression rules for recognizing dates and other time-denoting expressions.

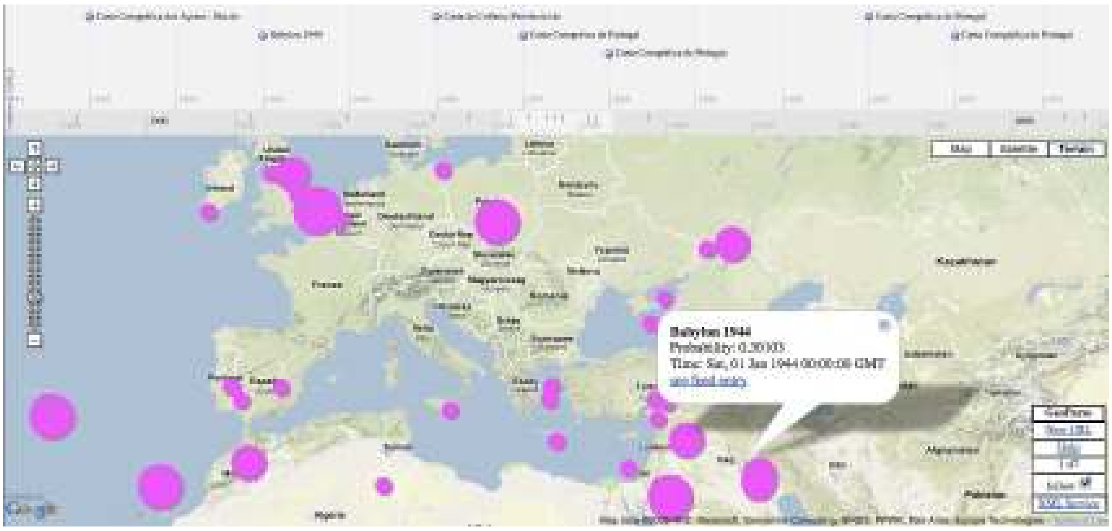
The dates recognized with regular expressions are converted into a canonical time-extent representation using rules. For the disambiguation of names of historical periods, we use queries to the gazetteer together with a simple set of heuristics. We start by making a simple gazetteer query for temporal features having the same exact name. A ranking score for each of the returned references is given as 1/n, where n is the number of returned references for that name. For the temporal references having more than one matching feature returned by the gazetteer query, we use related reference per discourse heuristics to try adjusting the ranking scores. Since names of time periods can vary according to the geographical location (e.g. *Revolution Period* can mean different things in different parts of the world), features having an association with a geographical feature that is also referenced in the document have their ranking score boosted by 0.2 up to a maximum value of 1. Features whose timeextent overlaps with that of other temporal features also referenced in the document have their score boosted by 0.2 up to a maximum value of 1.

For computing the temporal scope of the document we start by discarding all temporal references with a score bellow 0.5 (the score for the references extracted through rules is 1, as these are unambiguously assigned to a time-span). The scope is then given by the timespan that has a starting date corresponding to that of the earliest temporal reference in the document, and an ending date corresponding the latest temporal reference.

In the cases where the temporal information extraction fails to provide results, and if the RSS feed already defines a publication date for the items, we use this information to assign the temporal scope.

The geo-parser Web service developed in the context of DIGMAP

We developed a geo-parser Web service as part of the DIGMAP project, implementing the approaches described in the previous section for extracting geo-temporal context information from documents. The primary access interface for this geo-parser system is based on an XML format that resulted from extending OGC's proposal for a geo-parser service [4] in order to account with a) different types of geo-temporal references in the text, using GML to encode the geo-temporal information b) disambiguation scores associated with the discovered references, c) geo-temporal scopes assigned to the documents, and d) output format transformations through the use of XSLT filters.
Taking TimeMap[9] as inspiration, we also prototyped a simple exploratory user interface for showing the extracted geo-temporal information in a dual visualization, i.e. using maps and timelines – See Figure 3. The XSLT mechanism was used to transform the XML output into a suitable representation, i.e. one that could be interpreted by the Google Maps[10] and Simile Timeline JavaScript APIs[11]. An initial usability study with this prototype interface, involving five colleagues from our university departments, provided very interesting results. We are currently integrating mechanisms that use timelines and maps into the searching and browsing interfaces of the DIGMAP portal.



Babylon 1944
Probability: 0.30103
Time: Sat, 01 Jan 1944 00:00:00 GMT
see feed entry

Figure 3. A prototype user interface for geo-temporal exploration of information resources

Evaluation experiments

This section describes initial evaluation experiments performed with the proposed geo-parser service. The general evaluation methodology consisted of two steps. First, we measured the performance of the geo-parser service by simulating different workloads through the use of the Apache JMeter[12] tool. Next, we measured result quality by comparing the geographic and temporal footprints assigned to the textual resources against human-made annotations. The test collection used for this experiment consisted of a

[9] http://code.google.com/p/timemap/
[10] http://code.google.com/apis/maps/
[11] http://simile.mit.edu/timeline/
[12] http://jakarta.apache.org/jmeter/

set of metadata records from the DIGMAP catalogue, containing textual descriptions in multiple languages.

Evaluation of computational performance

With the Apache JMeter tool we simulated several simultaneous requests made to the geoparser service, measuring the response times in different conditions. This tool issues HTTP requests to the Web service and records the time to complete the requests. Both the service and the JMeter client were running on the same machine for our evaluations. This approach removed network latency, which can vary substantially. The experiments were performed on an Intel 2MGHz Core 2 Duo MacBook with 2GB RAM. The gazetteer and geo-parser services were both implemented as Java Web services, running on the same Apache Tomcat[13] application server. We ran a JMeter test corresponding to a maximum of 5 user request threads and with a ramp-up period of 2 seconds. The simulated requests were randomly chosen from a set of 100 examples involving the geo-parsing of text segments extracted from the Reuters-21578 corpus. Figure 4 presents the obtained results, showing that the prototype system can scale well to support demanding applications. Processing very large document collections should not be a difficult task to accomplish.

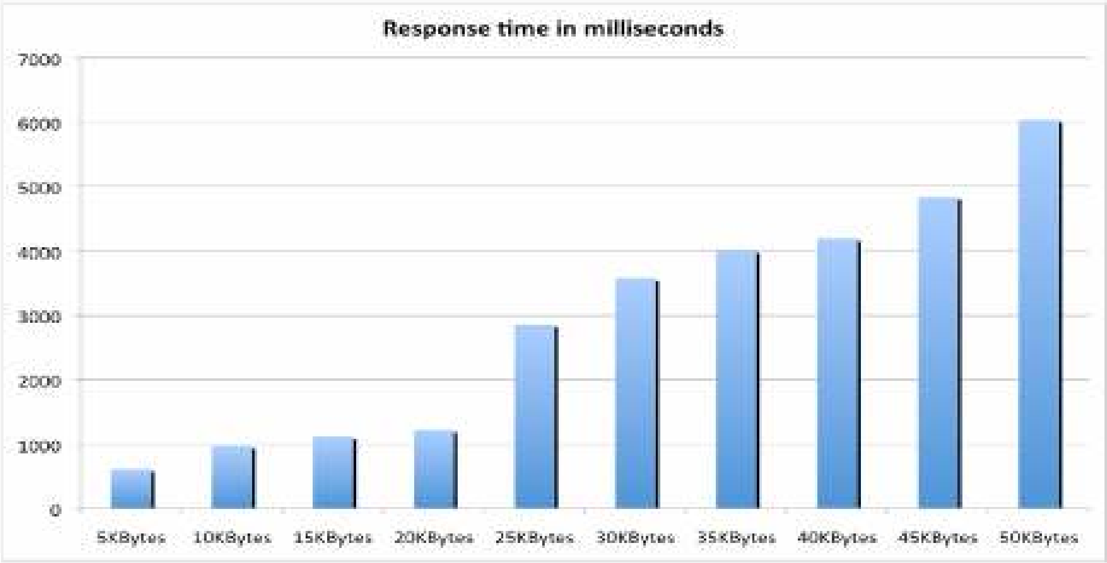


Figure 4. Results for experiments with JMeter

In the same test, we also measured the response times as a function of the size of the textual document given as input. The set of 100 examples was segmented according to size ranges, starting at 5Kb of text and moving up to 50Kb of text, with a step of 5. Figure 4 shows that the system scales almost linearly with the size of the input document. For 50Kb of text (above the average size of a Web page) and having 5 simultaneous requests, the system gives results in approximately six seconds (this includes the recognition of geo-temporal references in the text and the attribution of scopes). The obtained results can be improved through the use of caching mechanisms (in both the geo-parser and the gazetteer services) or by tuning the application server.

[13] http://tomcat.apache.org/

Evaluation of result quality

To evaluate the quality of the proposed extraction algorithms, we compare the geographical and temporal scopes that are automatically assigned to the documents against human-made annotations. The gold-standard collections of annotated textual resources consisted of a set of 511 metadata records from the DIGMAP digital library of old maps [7], having temporal and geographical annotations provided by librarians. The geographical scores are provided as minimum bounding boxes. The temporal scopes correspond to creation dates for the resources, sometimes given as a year date and others as a time period.

For the geographical scopes, we also compared the approach proposed in the paper against two simple baselines:

1. Generating the geographic scope with basis on the most frequent place reference
2. Generating the geographic scope with basis on the bounding box covering all the place references in the document with a score above a 0.5 threshold.

The considered evaluation metrics are based on the distance between the automatically assigned scopes and the human-made annotations:

- For the geographical annotations, and as a first measure, we used the ratio of the overlapping area between the scope in the annotation and automatically assigned scope (multiplied by two), and the sum of the areas for the two scopes.
- For the geographical annotations, we also used the distance between the centroid point of the geographic area automatically assigned as the scope and the centroid point for the geographic area that is provided in the annotation.
- For the temporal annotations, we use the difference in years between the middle dates of the time period assigned automatically and the period in the annotation.

By imposing thresholds on the metrics above (e.g. overlap ratio above 90%, distance bellow 25 kilometres and difference bellow 2 years) we can also measure results in terms of accuracy (i.e. the percentage of items assigned to correct scopes).

Table 2 presents the obtained results, showing that proposed method for assigning geographic scopes outperforms the simpler baselines used in our tests. Of the 511 resources used in the test, we could recognize place references in a significant percentage of them. In terms of accuracy, the results seem to be of sufficient quality for usage in a real-world application involving the use of the geographic document context. This is particularly true if we think of generalist and wide-coverage applications such as the DIGMAP portal, where data is mostly explored at the level of large geographical regions.

	Scopes Assigned	**Average Distance**	**Average Overlap**	**Accuracy 25 Km**	**Accuracy 100 Km**
Geographic	395 77%)	32 Km	0.82	0.56	0.81
Baseline 1	395 (77%)	38 Km	0.78	0.51	0.70
Baseline 2	392 (77%)	63 Km	0.67	0.31	0.52
Temporal	98 (19%)	8.7 Years			

Table 2. Results for the DIGMAP gold-standard collection.

It is interesting to note that the difference between the centroid of the assigned scopes and the centroid of the real scopes spans several Kilometres, although the results are still of acceptable quality for many applications (e.g. the assigned scopes have a high overlap with the scopes that were given in the annotations). Also regarding the area of overlap, it should be noted that the automatically assigned scopes were often given only as centroid coordinates, due to the fact that the gazetteer only contained this information. These cases were not accounted for in the computation of the average overlap.

In what concerns the recognition of temporal references, we encountered several problems. Most of the resources in the DIGMAP catalogue were only assigned to a year of creation, whereas the textual descriptions contained names of historical periods. Even when the descriptions contained the year of creation, it was given as a numeric value without further context in the text (we did not consider an extraction rule for these cases as it would also return many false positives corresponding to general numeric expressions). Currently ongoing work is addressing the evaluation of the temporal extraction procedures through the use of a more adequate document collection (i.e. a set of metadata records with richer descriptions), as well as the improvement of the extraction procedure.

Despite the problems with the temporal domain, the obtained results are encouraging and seem to be of sufficient quality for usage in many different applications. The DIGMAP project is currently exploring the usage of the geo-parser system for the automatic indexing of resources in its catalogue, for the parsing of search queries given in the DIGMAP portal, and for the construction of browsing interfaces that combine maps and timelines using information automatically extracted from the resources.

Conclusions and future work

Metadata records in digital libraries often describe resources that are relevant to somewhere at some particular time. Thus, these collections can be organized according to geographical and temporal criteria. Although the general idea seems simple, the resources are often characterized by textual descriptions and both place names and time periods are highly ambiguous (e.g. what is the meaning behind Gulf War Period or Lisbon). Name ambiguity must be resolved, in order to gain a full understanding of the involved geo-temporal context.

This paper argues that relatively simple extraction techniques can still provide results with a sufficiently high quality. This work also indicated that a correct interpretation of the temporal context of a document can be a harder task for automated methods than the interpretation of their geographic context. For future work, we will focus on improving results for the temporal domain. There are also many heuristics that can result in general improvements. Besides place types, place demographics or the document's language could be used in default sense heuristics. The thresholds involved in the proposed methods could also be target of further studies, in order to tune them to optimum values. Finally, the use of more advanced natural language processing methods (e.g. part-of-speech tagging) could also be attempted.

Acknowledgements

This research was partially funded by the eContentplus Programme of the European Community, under the contract ECP-2005-CULT-038042 (DIGMAP project).

References

[1] L. Hill and Q. Zheng 1999. Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. Proceedings of the American Society for Information Science Annual Meeting

[2] C. Jones, A. Abdelmoty, D. Finch, G. Fu and S. Vaid 2004 The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Proceedings of the 3rd International Conference on Geographic Information Science

[3] A. Kornai 2003 Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references

[4] J. Lansing 2001 Geoparser service draft candidate implementation specification. OGC Discussion Paper 01-035

[5] E. Amitay, N. Har'El, R. Sivan and A. Soffer 2004 Web-a-where: geotagging Web content. Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval

[6] H. Manguinhas, B. Martins, J. Borbinha 2008, The DIGMAP Web Gazetteer Service (to appear)

[7] B. Martins, J. Borbinha, G. Pedrosa, J. Gil and N. Freire 2007 Geographically-aware information retrieval for collections of digitized historical maps. Proceedings of the 4th Workshop on Geographical Information Retrieval

[8] B. Boguraev and R. K. Ando 2005 TimeML-compliant text analysis for temporal reasoning. Proceedings of the 19th International Joint Conference on Artificial Intelligence

[9] N. Chinchor 1998 Proceedings of the 7th Message Understanding Conference

[10] P. Clough and M. Sanderson 2004 A proposal for comparative evaluation of automatic annotation for georeferenced documents. Proceedings of the 1st Workshop on Geographic Information Retrieval

[11] W. Cohen and S. Sarawagi 2004 Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining

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

[13] E. Garbin and I. Mani 2005 Disambiguating toponyms in news. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing

[14] J. Pustejovsky, J. Castano, R. Ingria, R. Saurı, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev 2003 TimeML: Robust specification of event and temporal expressions in text. Proceedings of the AAAI Spring Symposium on New Directions in Question-Answering

[15] H. Li, K. R. Srihari, C. Niu and W. Li 2002 Location normalization for information extraction. Proceedings of the 19th Conference on Computational Linguistics

[16] R. Malouf 2002 Markov models for language-independent named entity recognition. In Proceedings of the 6th Conference on Natural Language Learning

[17] D. Manov, A. Kiryakov, B. Popov, K. Bontcheva, D. Maynard and H. Cunningham 2003 Experiments with geographic knowledge for information extraction. Proceedings of the HTL/NAACL-03 Workshop on Analysis of Geographic References

[18] B. Martins and M. J. Silva 2005 A graph-based ranking algorithm for geo-referencing documents. In Proceedings of the 5th IEEE International Conference on Data Mining

[19] A. Mikheev, M. Moens and C. Grover 1999 Named entity recognition without gazetteers. Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics

[20] E. T. K. Sang and F. De Meulder 2003 Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. Proceedings of the 7th Conference on Natural Language Learning

[21] Y. Chen, G. Di Fabbrizio, D. Gibbon, R. Jana, S. Jora, B. Renger and B. Wei 2007 GeoTracker: Geospatial and temporal RSS navigation. Proceedings of the 16th World Wide Web conference

[22] C. B. Jones and R. Purves 2006 GIR'05 : The 2005 ACM workshop on geographical information retrieval, ACM SIGIR Forum, 40(1)

[23] J. Leidner 2007 Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names, Ph.D. thesis, School of Informatics, University of Edinburgh, Scotland, UK

[24] M. Buckland and L. Lancaster 2004 Combining Place, Time, and Topic : The Electronic Cultural Atlas Initiative. D-Lib Magazine, 10(5)

[25] V. Petras, R. R. Larson and M. Buckland 2006 Time period directories: a metadata infrastructure for placing events in temporal and geographic context. Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries

[26] M. J. Bates and D. N. Wilde 1993 An analysis of search terminology used by humanities scholars: the Getty Online Searching Project Report Number 1. Library Quarterly, 63(1)

[27] M. Kimler 2004 Geo-coding: Recognition of geographical references in unstructured text, and their visualization. Diploma thesis, Fachhochschule Hof, Germany

[28] P. Harpring 1997 The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums, Archives and Museum Informatics

[29] W. Robert and S. Draper 1983 Questionnaires as a Software Evaluation Tool Interface Design. Proceedings of the ACM CHI Conference on Human Factors in Computing Systems