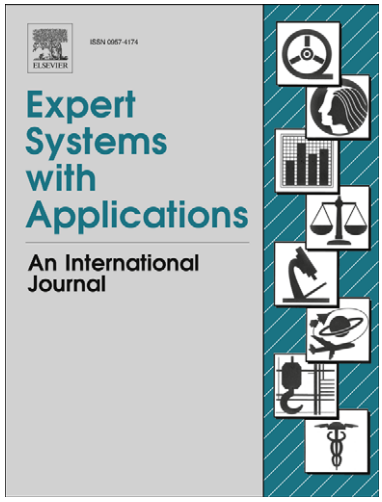

Application of Text Summarization techniques to the Geographical Information Retrieval task

José M. Perea-Ortega, Elena Lloret, L. Alfonso Ureña-López, Manuel Palomar

PII: S0957-4174(12)01253-5
DOI: http://dx.doi.org/10.1016/j.eswa.2012.12.012
Reference: ESWA 8262

To appear in: *Expert Systems with Applications*

Received Date: 20 October 2012
Accepted Date: 9 December 2012

Please cite this article as: Perea-Ortega, J.M., Lloret, E., Alfonso Ureña-López, L., Palomar, M., Application of Text Summarization techniques to the Geographical Information Retrieval task, *Expert Systems with Applications* (2012), doi: http://dx.doi.org/10.1016/j.eswa.2012.12.012

Application of Text Summarization techniques to the Geographical Information Retrieval task

José M. Perea-Ortega[a], Elena Lloret[b], L. Alfonso Ureña-López[a], Manuel Palomar[b]

[a]*SINAI research group. Computer Science Department*
University of Jaén, E-23071, Jaén, Spain
{*jmperea,laurena*}*@ujaen.es*
[b]*NLP research group. Department of Software and Computing Systems*
University of Alicante, E-03080, Alicante, Spain
{*elloret,mpalomar*}*@dlsi.ua.es*

Abstract

Automatic Text Summarization has been shown to be useful for Natural Language Processing tasks such as Question Answering or Text Classification and other related fields of computer science such as Information Retrieval. Since Geographical Information Retrieval can be considered as an extension of the Information Retrieval field, the generation of summaries could be integrated into these systems by acting as an intermediate stage, with the purpose of reducing the document length. In this manner, the access time for information searching will be improved, while at the same time relevant documents will be also retrieved. Therefore, in this paper we propose the generation of two types of summaries (generic and geographical) applying several compression rates in order to evaluate their effectiveness in the Geographical Information Retrieval task. The evaluation has been carried out using GeoCLEF as evaluation framework and following an Information Retrieval perspective without considering the geo-reranking phase commonly used in these systems. Although single-document summarization has not performed well in general, the slight improvements obtained for some types of the proposed summaries, particularly for those based on geographical information, made us believe that the integration of Text Summarization with Geographical Information Retrieval may be beneficial, and consequently, the experimental set-up developed in this research work serves as a basis for further investigations in this field.

Keywords: Geographical Information Retrieval, Text Summarization, Information Retrieval, GeoCLEF

1. Introduction

Nowadays geographic information is stored in a wide variety of media and types of documents. In recent decades, the technology used to access such information has focused on the combination of digital maps and databases, which is characteristic of most Geographic Information Systems (GIS). In fact, while in GIS users are interested in the extraction of information from a precise, structured, map-based representation, in Geographic Information Retrieval (GIR) users are interested in extracting information from unstructured textual information, by exploiting geographic references in queries and document collection to improve retrieval effectiveness. Only in recent years a lot of attention has been paid to the development of automatic systems that specifically deal with the retrieval of geographic information, available in the vast amount of unstructured documents in the Web. Therefore, GIR can be considered an active and growing research field concerned with improving the quality of geographically-specific Information Retrieval (IR), focusing on access to unstructured documents [11, 15]. In a GIR system, both queries and search documents are usually based on natural language, in contrast to the more formal approach common in GIS, where specific geo-referenced objects are retrieved from a structured database.

Another interesting research field related to IR is automatic Text Summarization (TS), whose aim is to keep the essential information of text documents by discarding the unnecessary details contained in them [30]. TS has been shown to be useful for other applications, such as IR, Question Answering (QA), or Text Classification (TC). Specifically, TS has never been applied to GIR before, which can be considered an extension of the IR field.

Therefore, the main research objective of this work is to analyze in detail the impact of the use of automatic summaries in the context of GIR systems. We hypothesize that a good summary or abstract should keep the main idea provided by the original document, and consequently, it could be integrated into GIR systems by acting as an intermediate stage. We focus on summarizing text documents that contain different locations or geographic places. In this way, the document size and their indexing time would be reduced and possible non-relevant passages could be discarded.

The remainder of this paper is structured as follows: in Section 2, the most important literature related to GIR and TS are expounded; in Section 3, we describe the GIR and TS systems used for the experiments carried out; in Section 4 and Section 5, the evaluation framework is briefly described and the experiments and results are presented, respectively; in Section 6, an analysis and discussion of the results is carried out. Finally in Section 7, the relevant conclusions are drawn and the future work is outlined.

2. Background

In this section some of the most important works related to both fields covered in this study are expounded: Geographic Information Retrieval (GIR) and Text Summarization (TS). Within the TS subsection, we briefly review different approaches that combine TS with IR.

2.1. Geographic Information Retrieval

According to Jones and Purves [11], GIR should provide facilities to retrieve and rank documents with respect to their relevance, among others, from an unstructured or partially structured collection on the basis of queries specifying both the theme and geographic scope. As a demonstration of the interest in GIR, some workshops and evaluation campaigns have been taking place since 2004. Last GIR'10 workshop[1] was held in cooperation with ACM SIGSPATIAL and it collects the latest research in the GIR field. GeoCLEF[2] [8] took place between 2005 and 2008 under the umbrella of the Cross Language Evaluation Forum (CLEF) providing a competitive framework in order to evaluate GIR systems. One of the main conclusions of GeoCLEF was that the addition of more geographic knowledge in GIR systems had a little effect on their performance. Finally, NTCIR GeoTime[3] is a recent track within NTCIR conferences that is focused on QA, taking into account both geographic and temporal constraints.

The architecture of a GIR system can be considered similar to that of the IR process. It is consists basically of two phases: indexing and searching. During both phases it is necessary to carry out text operations such as toponym detection and resolution. Toponym detection is usually performed by

[1] `http://www.geo.uzh.ch/~rsp/gir10`
[2] `http://ir.shef.ac.uk/geoclef`
[3] `http://metadata.berkeley.edu/NTCIR-GeoTime`

means of a Named Entity Recognizer (NER) module, while toponym disambiguation is a more complex task that has to take into account the context of the entity detected in order to establish its geographic scope. Both processes are usually supported by a Geographical Knowledge Base (GKB). GKB is a database that determines the connection from a name to a geographical entity and how two entities are connected between them. Gazetteers or geographical ontologies are examples of GKBs.

Regarding the indexing phase, two main indexes are generated: the spatial and text indexes. Spatial index includes all the geographic entities detected in each document and the text index incorporates the preprocessed terms obtained after applying the stemming process and removing the stopwords from each document. Both indexes will be used later during the search and reranking phases in order to retrieve the most relevant documents. Then, the preprocessed query is run against the search engine obtaining a list of relevant documents. Finally, the reranking process ranks again the documents retrieved according to a geographic relevance, representing one of the most important challenges for GIR systems.

The weighting scheme used is an important issue to consider in any IR system in general and in a GIR system in particular. The search engines used in GIR do not differ significantly from those used in standard IR. Gey et al. [8] noted that most GeoCLEF participants based their systems on the vector space model with TF·IDF weighting scheme. According to Perea-Ortega et al. [25], Terrier[4], Lemur[5] and Lucene[6] were the most used search engines by the GIR systems presented in GeoCLEF. These tools implement different weighting models by default, such as TF·IDF, BM25 or DFR.

In this study, we take into account only the indexing and searching phases of a GIR system, analyzing how the use of text summaries affects the system performance. For this reason the reranking process, which is the next step after document retrieval, is not addressed in this work. Therefore this analysis can be accomplished from the IR point of view exclusively, i.e., taking into account the indexing time and the average precision of the documents retrieved.

[4] http://terrier.org
[5] http://www.lemurproject.org
[6] http://lucene.apache.org

2.2. Text Summarization

The process of automatic TS mainly comprises three phases [10]: i) topic identification; ii) interpretation or topic fusion; and iii) summary generation.

The first phase (topic identification) consists of determining the particular subject of a document. It is usually approached by assigning a score to each unit (words, sentences, phrases, etc.), which is indicative of its importance. This is commonly done by means of machine learning algorithms [19, 32], statistical techniques [20, 23], discoursed-based approaches [4, 9], different types of linguistic knowledge, such as semantic [33], or by means of specialized resources, as for instance, Wikipedia [14]. In the end, the top score units up to a desired length are extracted.

The next phase is the interpretation or topic fusion. During this stage the topics identified as important are fused, represented in new terms, and expressed using a new formulation, which includes concepts or words not found in the original text. This stage is what distinguishes extractive [21] from abstractive summarization [7]. Finally, the summary generation only makes sense if abstractive summaries are generated; otherwise, the summary is a selection of sentences. In former cases, natural language generation techniques are needed to produce the final text of the summary.

Although research in TS has posed great interest for the community, the generated summaries are still far from ideal from a human point of view, partly due to the challenges associated to the generation of abstracts. However, it has been shown that automatic summaries, although imperfect in their nature, can be extremely useful for other applications, such as IR [12], QA [22], or TC [13]. In this manner, summaries can be integrated into these systems in order to reduce the size of the documents to be processed, keeping the essential information and removing the noisy one.

Focusing on IR, which is the scope of this paper, we can find previous approaches that combine IR and TS. The most common manner to combine both tasks is by taking the input for TS, the output of the IR system. In other words, the important documents are first retrieved, and then, a summary is generated, taking into account these documents. Therefore, IR helps to gather only relevant documents to a query, while TS selects the most important information from them. In light of this idea, Lin et al. [16] proposed an approach to identify and retrieve the most important concepts of a document that could be later exploit to generate a summary. Similarly, SWEeT [31] relies on a search engine to retrieve relevant documents to a query from the Web, and then summarization techniques based on Latent

Semantic Analysis (LSA) are used to identify and extract the most important sentences from the retrieved documents using, at the same time, cosine similarity to avoid redundancy in the final summaries. The QCS system [5] also integrates an IR module but, instead of retrieving documents directly from the Internet, it does so from a static document collection. Once the relevant documents have been retrieved, the system clusters them according to their main topic, and finally a summary is produced for each cluster.

On the contrary, less research has been carried out to analyze how text summaries can be beneficial for the IR process. Sakai and Spärck Jones [28] proved that generic summaries with a compression rate ranging from 10% to 30% were the most appropriate ones for the indexing stage in IR, concluding that a summary index was as effective as the full text index, for precision-oriented search. Furthermore, to the best of our knowledge, it has not been studied to what extent automatic summaries would be useful for more specific types of IR, such as GIR, taking them as part of the IR process, instead of using them for providing the output information in a more concise manner. Related to GIR, only in [1] and [2], TS was used for generating summaries about specific locations in order to provide users with more information when searching for a place. Again, summarization is applied once the retrieval has been performed. Therefore, in this research a detailed analysis of the capabilities of automatic summaries for the particular task of GIR is carried out. In this manner, TS techniques will be integrated as an intermediate stage in the whole process, to reduce the amount of text to be processed by GIR systems, thus shortening the indexing time as well.

3. System overview

The goal of this Section is to explain the individual systems involved in this study and how they have been used in combination. In this study, we analyze the performance of the SINAI-GIR system [27] after applying the TS techniques provided by COMPENDIUM summarizer [17], a modular tool to automatically generate text summaries. Two types of summaries are generated to be indexed by the GIR system. The first one is based on generic summaries, where we are more focused on extracting general but relevant information, whereas the second one is based on geographic summaries, which takes also into account the geographic entities that appear in the document. Moreover, several experiments regarding various summary lengths have been carried out. Then, the summaries were indexed and different geographic

queries were run against the GIR system in order to evaluate the proposed approach. To this end, we have taken as a basis GeoCLEF, as it is the most important evaluation framework in the GIR context.

3.1. SINAI-GIR: a modular GIR system

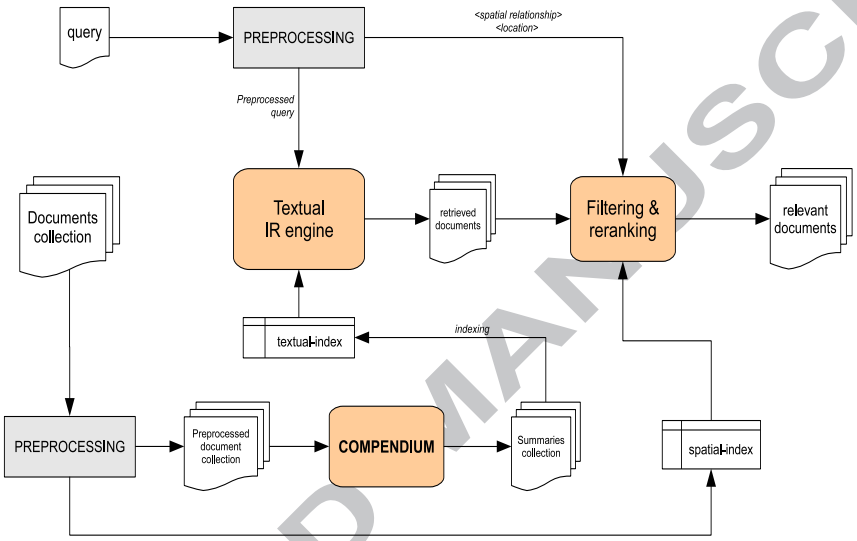

Figure 1: Overview of the SINAI-GIR system

The SINAI-GIR system has been used as a GIR system for the experiments carried out in this study. This system is composed of several modules as can be seen in Figure 1.

As regards document collection processing, it is based on detecting all the geographical entities in each document. With "*geographical entities*" or geographic information we refer exclusively to the place names contained in each document. We have incorporated the new module called COMPENDIUM responsible for generating the summaries from the original document collection. Then a spatial index is generated incorporating the place names detected for each document. During this preprocessing phase, the stopwords are also removed and the stems of each word are taken into account. We have used our own NER tool called GeoNER [26] to detect spatial entities. It is based on external knowledge resources such as GeoNames and Wikipedia.

Regarding query processing, each query is preprocessed and analyzed, identifying the geographical scope and the spatial relationship. To this end, we have used a Part Of Speech (POS) tagger like TreeTagger[7] along with some lexical and syntactic rules. Moreover, the stopwords are removed and the Snowball stemmer[8] is applied to each word of the query, except for the geographical entities.

During the text retrieval process we obtain 1,000 documents for each query. Due to the fact that SINAI-GIR is a modular system, it allows us to use different search engines applying therefore different weighting models. Specifically, we have used Terrier and Lemur in order to compare the behavior from different search engines. The weighting models used are *inL2* for Terrier and *BM25* for Lemur. As a final step, each preprocessed query (including their geographical entities) is run against the search engine.

It is important to note that, although GIR systems usually apply a geo-reranking process after the IR module, it is not particularly necessary in this study because we are interested in evaluating the Precision and Recall of the documents retrieved by the search engine using summaries instead of the original document collection.

3.2. COMPENDIUM: a modular tool to generate text summaries

COMPENDIUM is a modular text summarization tool that allows to produce a single-document summary[9] by means of different stages, as can be seen in Figure 2.

First of all, a basic *linguistic analysis* is applied to the input document, thus preparing it for further processing. To this end, external NLP state-of-the-art tools and resources are used. Specifically, this stage comprises sentence segmentation[10], tokenization[11], stemming[12], and stopword identification[13]. Then, for generating a summary, COMPENDIUM takes into con-

[7] TreeTagger v.3.2 for Linux. Available in `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[8] `http://snowball.tartarus.org`

[9] This means that the summary is generated taking as an input only one document.

[10] DUC sentence splitter: `http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz`

[11] Word Splitter: `http://cogcomp.cs.illinois.edu/page/tools_view/8`

[12] Porter Stemmer: `http://tartarus.org/~martin/PorterStemmer`

[13] English stopword list: `http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`

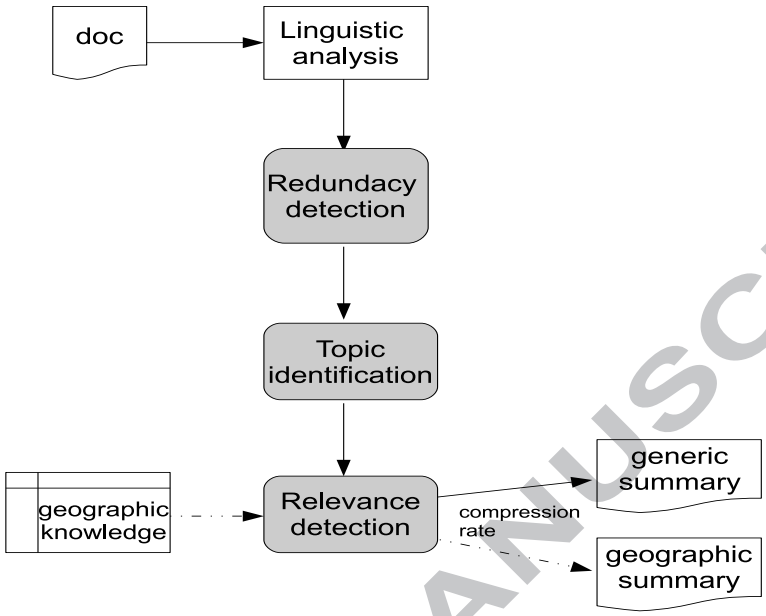

Figure 2: General architecture of COMPENDIUM

sideration two main issues: avoiding redundancy and identifying important content. For the former, the system detects repeated information by means of the *redundancy detection* stage, which employs textual entailment techniques [6] for determining when a given sentence has been already mentioned in the document. For the latter, COMPENDIUM rewards a sentence if it contains a larger number of noun-phrases where the elements have a high frequency within the document (without taking into account the stopwords). This is done is two steps: first the frequency of each term in the document is computed (we have called it *topic identification*), and then we compute the final relevance of a sentence based on the number and structure of its noun-phrases[14] (*relevance detection*). Combining both techniques, the weight of each word (i.e., its frequency computed in the *topic identification* stage) within a noun-phrase is counted, and then the score of the sentence is divided by the number of noun-phrases the sentence contains, as follows:

[14]BaseNP Chunker was used for extracting noun-phrases: `ftp://ftp.cis.upenn.edu/pub/chunker`

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \quad (1)$$

where:

r_{s_i} = is the relevance of sentence i,
$\#NPi$ = number of noun-phrases contained in sentence i,
tf_w = frequency of word w that belongs to the sentence's noun-phrase.

After applying these two intermediate stages, the last one is to produce the summary, guaranteeing that it has the desired length. Thus, COMPENDIUM produces generic informative summaries following an extractive paradigm and preserving the same order as the sentences had in the original document. In this manner it would not be too detrimental to the coherence of the generated summary.

The adaptation of COMPENDIUM for producing a new type of summary would be very easy, due to the modularity of its architecture. Concerning this, besides producing generic summaries, we adapt the summarizer in order to generate geographic summaries, so we can also quantify the impact of this kind of summaries for GIR, since they may contain not only the most relevant information of a document, but also the specific one related to a geographic entity. To this end, we introduce a new intermediate stage within the identification of important content, where the knowledge of the SINAI-GIR system about the geographical entities is also taken into account by COMPENDIUM for assigning more relevance to those sentences containing such spatial entities. Specifically, COMPENDIUM used the geographical entities recognized by the SINAI-GIR system during the document collection processing.

4. Evaluation framework

In order to evaluate the proposed approach we have used the GeoCLEF framework as it was already mentioned. GeoCLEF provides a document collection that consists of 169,477 documents composed of stories and newswires from the British newspaper *Glasgow Herald* (1995) and the American newspaper *Los Angeles Times* (1994), representing a wide variety of geographical regions and places. GeoCLEF also provides a total of 100 textual queries (25 per year). These are composed of three main fields: *title* (T), *description* (D) and *narrative* (N). Figure 3 shows an example of a query from GeoCLEF. For the experiments carried out in this study, we have employed the 100

queries provided by the GeoCLEF organizers during the four editions (from 2005 to 2008). Specifically, we have only taken into account the *title* field of them because it represents in a similar way how a user would launch a geographic query to a search engine. Some examples of the *title* field used in GeoCLEF queries are: "*vegetable exporters of Europe*", "*forest fires in north of Portugal*", "*airplane crashes close to Russian cities*" or "*natural disasters in the Western USA*".

```
<top>
  <num>10.2452/58-GC</num>
  <title>Travel problems at major airports near to London</title>
  <desc>To be relevant, documents must describe travel problems
  at one of the major airports close to London.</desc>
  <narr>Major airports to be listed include Heathrow, Gatwick,
  Luton, Stanstead and London City airport.</narr>
</top>
```

Figure 3: Example of a GeoCLEF query

A classification of the GeoCLEF queries depending on their geographic constraint was presented by Overell [24]. This classification is shown in Table 1.

Regarding the evaluation measures employed, results have been evaluated using the relevance judgements provided by the GeoCLEF organizers and the TREC evaluation method. The evaluation has been accomplished by using the Mean Average Precision (MAP) and Recall (R). The MAP measure computes the average precision over all queries. The average precision is defined as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved. Recall is a measure of the extent to which relevant documents are found or retrieved. Recall is 1.0 when every relevant document is retrieved.

5. Experiments and results

Different experiments have been carried out in order to evaluate the performance of the use of text summaries instead of the original documents of the GeoCLEF collection. Moreover, two different search tools (Terrier and Lemur) were used within the SINAI-GIR system applying distinct weighting models. As mentioned above, no spatial reranking is applied after the

Freq.	Geographic scope	Query example
9	Scotland	Walking holidays in Scotland
1	California	Shark Attacks off Australia and California
3	USA (excluding California)	Scientific research in New England Universities
7	UK (excluding Scotland)	Roman cities in the UK and Germany
46	Europe (excluding the UK)	Trade Unions in Europe
16	Asia	Solar or lunar eclipse in Southeast Asia
7	Africa	Diamond trade in Angola and South Africa
1	Australasia	Shark Attacks off Australia and California
3	North America (excluding the USA)	Fishing in Newfoundland and Greenland
2	South America	Tourism in Northeast Brazil
8	Other Specific Region	Shipwrecks in the Atlantic Ocean
6	Other	Beaches with sharks

Table 1: Classification of GeoCLEF queries according to their geographic constraint Overell [24]

searching so we use the default ranking provided by the search engine for each query.

Two types of summaries were generated using the COMPENDIUM tool:

- ***Generic* summaries**. The aim of these summaries is to provide a general overview of the main contents of a document. In this case, we are interested in checking whether this information would be enough to answer a complex query.

- ***Geographic* summaries**. These summaries take into account and assign more relevance to those sentences containing geographic entities. In this case, we first obtain all the entities recognized by the SINAI-GIR system during the document collection processing. Then we discard those sentences that not contain any spatial entity and for the remaining sentences we generate the summary of the document.

When generating a summary, it is important to take into consideration an appropriate size. This issue depends on different factors, such as the purpose of the summary, its informativeness, or the interests of the user. When the

ideal length for a summary is not known a priori, different compression rates should be analyzed in order to determine which would be the best. The compression rate can be defined as the how much shorter the summary is with respect to the original document (see Formula 2):

$$Compression\ rate = \frac{lengthS}{lengthD} \quad (2)$$

where S is the summary and D is the original document.

Compression rate can be computed according to different granularities, e.g., the number of words or the number of sentences. In our case, it indicates the proportion of sentences that are kept in the summary with respect to the total number of document sentences. Therefore, a compression rate of 40% means that the summary would contain only 40% of the sentences (e.g., if the document has 150 sentences, the summary would have only 60). Figure 4 illustrates how different compression rates would be obtained for an automatic summary, taking into account the number of sentences. Specifically for the experiments carried out in this study we generated summaries for compression rates ranging from 20% to 90%, with an increment of 20%. We observed that the proposed compression rates were enough to carry out a deep analysis of the impact of automatic summaries for the GIR task.

Below we show the results obtained from an IR perspective (using the MAP and Recall metrics) and from the point of view of the time employed during the indexing phase.

5.1. Evaluation from an IR perspective

Tables 2, 3, 4 and 5 show the MAP and Recall scores obtained for the 2005, 2006, 2007 and 2008 GeoCLEF editions, respectively. These results are compared with the baseline results obtained using the original document collection.

Analyzing these results from an IR perspective, we can observe that the summaries generated do not improve the baseline results using the 2005 and 2006 queries. Only using 2005 queries and Terrier as search engine, the generic summaries with a compression rate of 90% get the same Recall as the base case (0.8803). However, when the 2007 and 2008 queries are employed we obtain better results using the summaries rather than the original document collection in some cases. For example if we make use of the geographic summaries with a compression rate of 60% and Lemur as search engine, we

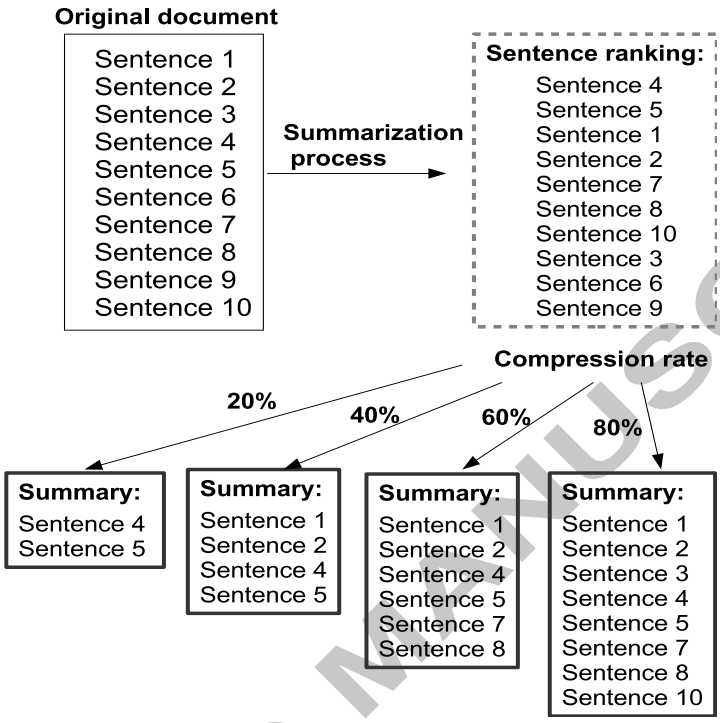

Figure 4: Illustrative explanation of the compression rate process

Comp. rate	Summary type	Terrier		Lemur	
		MAP	R	MAP	R
20%	generic	0.2104	0.5953	0.1843	0.6206
	geographic	0.3319	0.7821	0.2982	0.7791
40%	generic	0.2758	0.7363	0.2519	0.7276
	geographic	0.3442	0.8180	0.3072	0.7937
60%	generic	0.3191	0.8064	0.3006	0.8064
	geographic	0.3603	0.8531	0.3130	0.8210
80%	generic	0.3599	0.8540	0.3091	0.8307
	geographic	0.3604	0.8647	0.3113	0.8268
90%	generic	0.3688	**0.8803**	0.3297	0.8657
	geographic	0.3653	0.8793	0.3183	0.8570
Baseline (no summaries)		**0.3868**	**0.8803**	**0.3388**	**0.8696**

Table 2: Results obtained using the GeoCLEF 2005 queries

Comp. rate	Summary type	Terrier		Lemur	
		MAP	R	MAP	R
20%	generic	0.1401	0.5449	0.1368	0.5476
	geographic	0.2407	0.7698	0.1918	0.7248
40%	generic	0.1592	0.6137	0.1505	0.5952
	geographic	0.2298	0.7857	0.1963	0.7328
60%	generic	0.2001	0.7486	0.1910	0.7169
	geographic	0.2282	0.7962	0.1871	0.7380
80%	generic	0.2216	0.7248	0.1950	0.6798
	geographic	0.2278	0.7962	0.1937	0.7566
90%	generic	0.2272	0.7671	0.1963	0.7619
	geographic	0.2280	0.7936	0.2021	0.7619
Baseline (no summaries)		**0.2535**	**0.8148**	**0.2026**	**0.7671**

Table 3: Results obtained using the GeoCLEF 2006 queries

Comp. rate	Summary type	Terrier		Lemur	
		MAP	R	MAP	R
20%	generic	0.1007	0.5723	0.0973	0.6092
	geographic	0.2064	0.8000	0.1927	0.7938
40%	generic	0.1726	0.6846	0.1722	0.7000
	geographic	0.2326	0.8076	0.1966	0.8061
60%	generic	0.2075	0.8030	0.1954	0.7876
	geographic	0.2219	0.8492	**0.2161**	**0.8261**
80%	generic	0.2295	0.8107	**0.2116**	0.8000
	geographic	0.2279	0.8446	**0.2049**	0.8184
90%	generic	0.2268	0.8107	0.1854	0.7753
	geographic	0.2312	0.8584	**0.2014**	0.8138
Baseline (no summaries)		**0.2585**	**0.8769**	0.2005	0.8200

Table 4: Results obtained using the GeoCLEF 2007 queries

Comp. rate	Summary type	Terrier		Lemur	
		MAP	R	MAP	R
20%	generic	0.1702	0.4979	0.1385	0.5073
	geographic	0.2458	0.6559	0.1903	0.6894
40%	generic	0.1605	0.5676	0.1496	0.6425
	geographic	0.2604	0.6840	0.1989	0.7135
60%	generic	0.2471	0.6706	0.1874	0.6626
	geographic	0.2712	0.6974	**0.2199**	0.7175
80%	generic	0.2662	0.6720	0.2046	0.6599
	geographic	0.2644	0.6974	**0.2179**	0.7148
90%	generic	**0.2773**	0.6746	0.2143	0.7028
	geographic	0.2687	0.7068	0.2134	0.7202
Baseline (no summaries)		0.2713	**0.7242**	0.2149	**0.7523**

Table 5: Results obtained using the GeoCLEF 2008 queries

obtain an improvement of 7.78% of MAP score regarding the base case using the 2007 queries. Moreover, the MAP scores obtained using the geographic summaries with compression rates of 80% and 90% and the generic summaries with a compression rate of 80% for the same set of queries and search engine also outperform the base case with an improvement of 2.19%, 0.45% and 5.54%, respectively. Finally, the geographic summaries with compression rates of 60% and 80% improve the MAP score of the base case, as well, when Lemur is used as search engine for the 2008 queries. These improvements are 2.33% and 1.40%, respectively. Using the same set of queries but the other search tool (Terrier), the generic summaries with a compression rate of 90% also outperform the baseline MAP score with an improvement of 2.21%.

5.2. Evaluation of the indexing time vs. average MAP score

The usefulness of automatic TS in the GIR task can be also studied from the point of view of the indexing time. When the search collection is composed of a huge number of documents, reducing the indexing time would improve significantly the overall system performance. For this reason we have also measured the time taken by the search engine in order to index each type of summary generated, comparing it with the indexing time obtained using the original document collection. Concerning this time, we only took into account the indexing time once the summaries were already generated. The reason why not considering the whole process together is due to the fact that

in this research we want to analyze whether automatic summaries could have a positive impact in GIR systems. Figure 5 shows the comparison between the indexing time and the average MAP scores for each type of summary and each search engine.

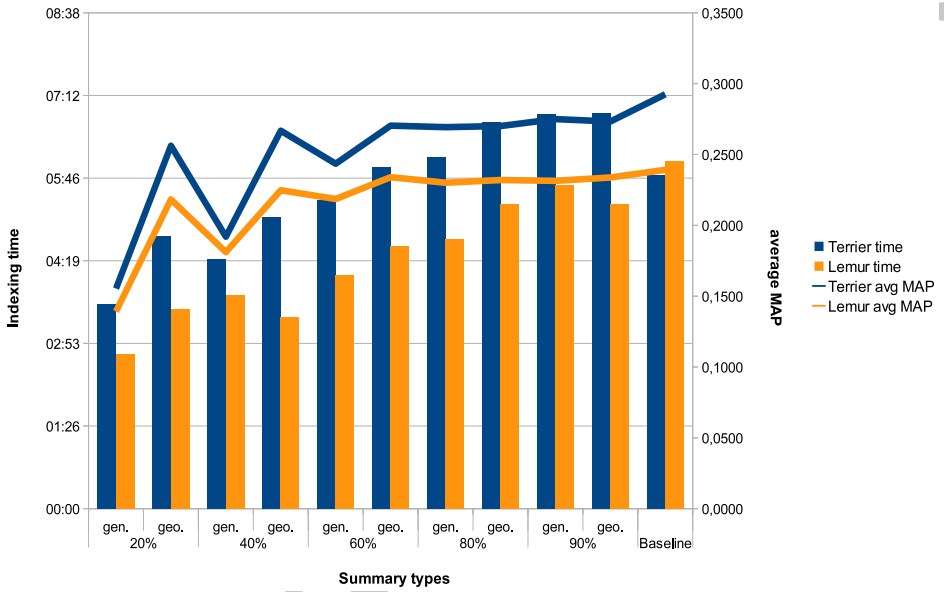

Figure 5: Indexing times vs. average MAP comparison

We can observe in Figure 5 that Lemur is faster than Terrier for all the compression rates tested. However for the base case this difference is not significant, since the indexing time achieved by Terrier is 14 seconds faster than that obtained by Lemur. It is curious the behavior of Terrier for the generic summaries with a compression rate of 80% and 90% because these outperform the indexing time for the base case (18 and 63 seconds, respectively). For these cases, geographic summaries usually spent more indexing time than generic summaries. On the other hand, as expected, there is a general increase of the indexing time when the compression rate grows.

Considering these results in conjunction with those regarding the evaluation from an IR perspective, geographic summaries of a 60% compression rate seemed the most appropriate ones, specially for geographic summaries.

However, it is important to stress the good performance of the geographic summaries with 20% and 40% of compression rate because their average MAP scores are slightly worse than the best ones but their indexing times are quite lower than those obtained by the best summaries regarding the average MAP score.

6. Analysis and discussion

From the results obtained, it can be noticed that we are facing a non-trivial task, and its difficulty has increased along the years. As can be seen, the results for the latest GeoCLEF editions (e.g., 2007 and 2008), both for the base case and the summaries, are generally lower than those held in 2005 and 2006.

In order to verify if the improvements obtained using summaries were statistically significant, a significance test was also carried out. We applied the Fisher's randomization test [3] only on those experiments in which improvements were obtained regarding the base case. According to Smucker et al. [29], who evaluated different tests of statistical significance for a large collection of TREC ad-hoc retrieval system pairs, the randomization test is recommended to evaluate IR systems. Therefore, we applied this test using the MAP scores obtained for each of the 100 queries only for the best experiments. Table 6 shows these results.

Best experiments	**MAP**	***p-value***
Lemur-60%-geographic	0.2161 (2007 queries) 0.2199 (2008 queries)	0.40022
Lemur-80%-generic	0.2116 (2007 queries)	0.19815
Lemur-80%-geographic	0.2049 (2007 queries) 0.2179 (2008 queries)	0.20584
Lemur-90%-geographic	0.2014 (2007 queries)	0.18966
Terrier-90%-generic	0.2773 (2008 queries)	0.01916

Table 6: Results of the Fisher's randomize tests for the best experiments

As can be seen in Table 6, only for the experiment "*Terrier-90%-generic*" we obtained a *p-value* less than 0.05, so the improvement achieved with that experiment can be considered statistically significant.

Contrary to our expectations, the results obtained when integrating TS into a GIR system were lower than expected. The reason why these cannot

generally outperform the base case may be due to the fact that the GeoCLEF queries require some general information that cannot be addressed with the summarization techniques proposed. Despite the fact that extractive statistical single-document summarization techniques have been shown to be useful for more specific and focused tasks, such as QA [18], the type of queries we are dealing with involve much more information. Analysing in depth the type of questions and the documents in the collection, we found that along the different GeoCLEF editions, the questions became more and more complex. For instance, for the following GeoCLEF query "*Golf tournaments in Europe. About golf tournaments held in European locations*", require broader information with a deeper semantic analysis of the text. For instance, if a document contains the entity *"Scotland"*, the sytem would have to infere that this place is in Europe, and therefore, sentences containing *"Scotland"* would be also rewarded more relevance. In contrast, this would be difficult to capture the whole information for the query by using extractive statistical single-document summaries, as in our case, since the generated summary will limit to the information contained in the document, using the same vocabulary.

Focusing on the two types of summaries generated, geographic summaries have a better performance than generic ones, as expected. It is noteworthy that in almost all cases (for different compression rates) the MAP score obtained using geographic summaries outperforms that obtained using generic summaries because GeoCLEF is an IR task in which geographic information plays a key role. Additionally, we would like to stress the fact that especially for higher compression rates (e.g., 20%) the increase of performance for geographic summaries compared to the generic ones is remarkable (e.g., 0.5449 vs. 0.7698 for generic and geographic summaries of 20% compression rate, according to the Recall value for the GeoCLEF 2006 queries using Terrier as search engine). The main reason for this behavior is that geographic summaries take into consideration the spatial entities found in the original documents so it is more likely that GIR system finds relevant document in this type of summary when geographic queries are used.

Regarding the different compression rates employed in order to generate the summaries, it is clear that the summaries with compression rates between 20% and 40% should not be used in this framework. This poor performance is due to the fact that these levels of compression are very restrictive, and as a consequence, other relevant sentences as well may be discarded. However, when using summaries with compression rates of 60% and 80% we can obtain

better results than base cases, as has been mentioned above. Therefore the use of this type of summaries can be an interesting strategy to apply in the GIR task.

The experimental set-up analyzed in this study has laid the foundations for further investigations, so the next step would be to analyzed and employed other techniques for producing summaries in order to improve the search time and precision of the retrieved documents. We strongly believe that the addtion of semantic knowledge, as well as the use of abstractive techniques would be more appropriate, since this manner the information could be generalized and the summaries could be tailored and adapted to each query.

7. Conclusions and future work

In this paper we analyzed the appropriateness of automatic summaries in the context of Geographical Information Retrieval (GIR). In particular, we proposed the generation of extractive statistical single-document summaries to act as an intermediate stage for the GIR task. Using this approach, the size of the document collection is reduced by keeping the essential information of the search documents.

Two types of summaries have been generated in this work: generic summaries, whose aim is to provide a general overview of the main contents of a document, and geographic summaries, that assign more relevance to those sentences containing geographic entities. Moreover, different compression rates have been tested during the experiments, ranging from 20% to 90%, with an increment of 20%. We have studied the performance of this approach from two points of view: evaluating the effectiveness of the IR process without considering the reranking phase commonly used in GIR systems, and evaluating the indexing time for two different search engines employed in the GIR system. To carry out this evaluation we have used GeoCLEF as framework, showing that the use of both types of summaries might improve the effectiveness of the IR process in some cases.

As main conclusion, the novelty of integrating automatic summarization techniques into the GIR process has been proven to obtain slight improvements for some types of the proposed summaries, particularly for those based on geographical information which took into account the geographic entities detected in the document collection. This issue may be of great interest for the GIR research community, and the experimental framework developed can

serve as a basis for further comparisons and analysis. Despite of this, the proposed approach based on the use of extractive statistical single-document summarization is not sufficient, stressing the need for experimenting with other types of summaries.

As future work, we will address several issues in two directions. On the one hand, we will analyze more thoroughly for what type of geographic queries the use of summaries achieves better performance. Moreover, we will apply the next step in our GIR architecture after the IR process, the georeranking process, in order to determine what retrieved documents should be set in the first positions according to the summaries indexed. On the other hand, we will analyze whether the addition of semantic knowledge and the use of abstractive summarization techniques lead to better results when integrating summaries within GIR systems. We will also propose other TS techniques in order not to be so restrictive when producing geographic summaries. From our experiments, we have found that by discarding all sentences that do not contain any geographic information may lead to a loss of information, since there may exist links between sentences. Therefore, we will analyse this issue in detail, by studying graph based algorithms that capture the relationship between sentences.

Acknowledgements

This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. It has also been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), projects TEXT-MESS 2.0 (TIN2009-13391-C04-01) and TEXT-COOL 2.0 (TIN2009-13391-C04-02) from the Spanish Government, a grant from the Valencian Government, project "Desarrollo de Técnicas Inteligentes e Interactivas de Minería de Textos" (PROMETEO/2009/119), and a grant no. ACOMP/2011/001.

References

[1] Cai, C., Hovy, E., 2010. Summarizing textual information about locations in a geo-spatial information display system. In: Proceedings of the

NAACL HLT 2010 Demonstration Session. Association for Computational Linguistics, Los Angeles, California, pp. 5–8.

[2] Cai, C., Hovy, E., 2011. Summarizing textual information about locations. In: Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications. ACM, New York, NY, USA, pp. 1–9.

[3] Cohen, P. R., 1995. Empirical methods for artificial intelligence. MIT Press.

[4] Cristea, D., Postolache, O., Pistol, I., 2005. Summarisation through discourse structure. In: Proceedings of the Computational Linguistics and Intelligent Text Processing, 6th International Conference. pp. 632–644.

[5] Dunlavy, D. M., O'Leary, D. P., Conroy, J. M., Schlesinger, J. D., 2007. QCS: A System for Querying, Clustering and Summarizing Documents. Information Processing & Management 43 (6), 1588–1605.

[6] Ferrández, O., 2009. Textual entailment recognition and its applicability in nlp task. Ph.D. thesis, Universidad de Alicante.

[7] Genest, P.-E., Lapalme, G., 2011. Framework for abstractive summarization using text-to-text generation. In: Proceedings of the Workshop on Monolingual Text-To-Text Generation. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 64–73.

[8] Gey, F. C., Larson, R. R., Sanderson, M., Joho, H., Clough, P., Petras, V., 2005. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: CLEF. Vol. 4022 of Lecture Notes in Computer Science. Springer, pp. 908–919.

[9] Gonçalves, P. N., Rino, L., Vieira, R., 2008. Summarizing and Referring: Towards Cohesive Extracts. In: DocEng '08: Proceeding of the 8th ACM Symposium on Document Engineering. pp. 253–256.

[10] Hovy, E., 2005. The Oxford Handbook of Computational Linguistics. Oxford University Press, Ch. Text Summarization, pp. 583–598.

[11] Jones, C. B., Purves, R. S., 2008. Geographical information retrieval. International Journal of Geographical Information Science 22 (3), 219–228.

[12] Kan, M.-Y., Klavans, J. L., 2002. Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries. pp. 36–45.

[13] Ker, S. J., Chen, J.-N., 2000. A Text Categorization based on Summarization Technique. In: Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval. pp. 79–83.

[14] Kumar, N., Srinathan, K., Varma, V., 2012. Using wikipedia anchor text and weighted clustering coefficient to enhance the traditional multi-document summarization. In: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing. Vol. 7182 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 390–401.

[15] Larson, R., 1996. Geographic information retrieval and spatial browsing. In: Smith, Gluck, M. (Eds.), Geographic Information Systems and Libraries: Patronsand Mapsand and Spatial Information. pp. 81–124.

[16] Lin, H.-T., Chi, N.-W., Hsieh, S.-H., 2012. A concept-based information retrieval approach for engineering domain-specific technical documents. Advanced Engineering Informatics 26 (2), 349 – 360.

[17] Lloret, E., 2011. Text summarisation based on human language technologies and its applications. Ph.D. thesis, University of Alicante.

[18] Lloret, E., Llorens, H., Moreda, P., Saquete, E., Palomar, M., Dec. 2011. Text summarization contribution to semantic question answering: New approaches for finding answers on the web. International Journal of Intelligent Systems 26 (12), 1125–1152.

[19] Manne, S., Shaik Mohd., Z., Sameen Fatima, S., 2012. Extraction based automatic text summarization system with HMM tagger. In: Satapathy, S., Avadhani, P., Abraham, A. (Eds.), Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. Vol.

132 of Advances in Intelligent and Soft Computing. Springer Berlin / Heidelberg, pp. 421–428.

[20] McCargar, V., 2005. Statistical approaches to automatic text summarization. Bulletin of the American Society for Information Science and Technology 30 (4), 21–25.

[21] Mei, J.-P., Chen, L., 2012. SumCR: A new subtopic-based extractive approach for text summarization. Knowledge and Information Systems 31, 527–545.

[22] Mori, T., Nozawa, M., Asada, Y., 2004. Multi-answer-focused Multi-document Summarization using a Question-answering Engine. In: Proceedings of the 20th International Conference on Computational Linguistics. pp. 439–445.

[23] Orăsan, C., 2009. Comparative Evaluation of Term-Weighting Methods for Automatic Summarization. Journal of Quantitative Linguistics 16 (1), 67–95.

[24] Overell, S. E., 2009. Geographic information retrieval: Classification, disambiguation and modelling. Master's thesis, Department of Computing, Imperial College.

[25] Perea-Ortega, J. M., García-Cumbreras, M. A., García-Vega, M., Ureña-López, L. A., 2008. Comparing several textual information retrieval systems for the geographical information retrieval task. In: Kapetanios, E., Sugumaran, V., Spiliopoulou, M. (Eds.), Natural Language and Information Systems. Vol. 5039 of Lecture Notes in Computer Science. Springer, pp. 142–147.

[26] Perea-Ortega, J. M., Martínez-Santiago, F., Montejo-Ráez, A., Ureña-López, L. A., 2009. Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. Procesamiento del Lenguaje Natural 43, 33–40.

[27] Perea-Ortega, J. M., Ureña-López, L. A., García-Vega, M., García-Cumbreras, M. A., 2008. Using query reformulation and keywords in the geographic information retrieval task. In: CLEF. Vol. 5706 of Lecture Notes in Computer Science. Springer, pp. 855–862.

[28] Sakai, T., Spärck Jones, K., 2001. Generic Summaries for Indexing in Information Retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 190–198.

[29] Smucker, M. D., Allan, J., Carterette, B., 2007. A comparison of statistical significance tests for information retrieval evaluation. In: CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, New York, NY, USA, pp. 623–632.

[30] Spärck Jones, K., 2007. Automatic summarising: The State of the Art. Information Processing & Management 43 (6), 1449–1481.

[31] Steinberger, J., Jezek, K., Sloup, M., 2008. Web topic summarization. In: Proceedings of the 12th International Conference on Electronic Publishing. pp. 322–334.

[32] Wong, T.-L., Lam, W., 2008. Learning to extract and summarize hot item features from multiple auction web sites. Knowledge and Information Systems 14, 143–160.

[33] Zhu, X., Ding, W., Yu, P., Zhang, C., 2011. One-class learning and concept summarization for data streams. Knowledge and Information Systems 28, 523–553.