TALP at GikiCLEF 2009

Daniel Ferrés and Horacio Rodríguez
TALP Research Center
Software Department
Universitat Politècnica de Catalunya
{*dferres,horacio*}*@lsi.upc.edu*

Abstract

This paper describes our experiments in Geographical Information Retrieval with the Wikipedia collection in the context of our participation in the GikiCLEF 2009 Multilingual task in English and Spanish. Our system, called gikiTALP, follows a very simple approach that uses standard Information Retrieval with the Sphinx full-text search engine and some Natural Language Processing techniques without Geographical Knowdledge.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Design, Performance, Experimentation

Keywords

report. Information Retrieval, Wikipedia, Geographical Information Retrieval, Natural Language Processing

1 Introduction

In this paper we present the overall architecture of our gikiTALP IR system and we describe briefly its main components. We also present the experiments, results and initial conclusions in the context of the GikiCLEF 2009 Monolingual English ans Spanish task.

GikiCLEF 2009 is an evaluation task under the scope of CLEF. Its aim is to evaluate systems which find Wikipedia entries/documents that answer a particular information need, which requires geographical reasoning of some sort. GikiCLEF is the successor of the GikiP 2008 [3] pilot task which ran in 2008 under GeoCLEF.

For GikiCLEF, systems will need to answer or address geographically challenging topics, on the Wikipedia collections, returning Wikipedia document titles as list of answers in all languages it can find answers.

The following (Wikipedia) languages are available in GikiCLEF: Bulgarian, Dutch, English, German, Italian, Norwegian, Portuguese, Romanian and Spanish.

1.1 GikiCLEF collections

The Wikipedia collections for all GikiCLEF languages are available in three formats, HTML dump, SQL dump, and XML version. Most of the collections are from June 2008. We used the SQL dump version of the English and Spanish collections.

Table 1: Description of the Collections we used at gikiclef 2009.

Language	#Total	#Pages	#Templates	#Categories	#Images
en	6,587,912	5,255,077	154,788	365,210	812,837
es	714,294	641,852	11,885	60,556	1

2 System Description

The system architecture has three phases that are performed sequentially: Collection Indexing, Topic Analysis, and Information Retrieval. The textual Collection Indexing has been applied over the textual collections with MySQL and the full-text engine Sphinx using the Wikipedia SQL dumps.

Sphinx [1] is a full-text search engine that provides fast, size-efficient and relevant full-text search functions to other applications. The indexes created with Sphinx do not have any language processing. Sphinx has two types of weighting functions:

- Phrase rank: based on a length of longest common subsequence (LCS) of search words between document body and query phrase.
- Statistical rank: based on classic BM25 function which only takes word frequencies into account.

We used two types of search modes in Sphinx:

- MATCH ALL: the final weight is a sum of weighted phrase ranks.
- MATCH EXTENDED: the final weight is a sum of weighted phrase ranks and BM25 weight, multiplied by 1000 and rounded to integer.

The Topic Analysis phase extracts some relevant keywords (with its analysis) from the topics. These keywords are then used by the Document Retrieval phases. This process extracts lexico-semantic information using the following set of Natural Language Processing tools: **TnT** (POS tagger) and [2] **WordNet lemmatizer** (version 2.0) for English, and **Freeling** [1]. for Spanish.

The retrieval is done with Sphinx and then the final results are filtered. The Wikipedia entries without Categories are discarded.

3 Experiments

For the GikiCLEF 2009 evaluation we designed a set of three experiments that consist in applying different baseline configurations (see Table 2) to retrieve Wikipedia entries (answers) of 50 geographically challenging topics.

The three baseline runs were designed changing two parameters of the system: the IR Sphinx search mode and the Natural Language Processing techniques applied over the query. The first run (gikiTALP1) do not uses any NLP processing technique over the query and the Sphinx match mode used is MATCH_ALL. The second run (gikiTALP2) uses stopwords filtering and the lemmas of the remaining words as a query and the Sphinx match mode used is MATCH_ALL. The third run (gikiTALP3) uses stopwords filtering and the lemmas of the remaining words as a query and the Sphinx match mode used is MATCH_EXTENDED.

[1] `http://www.sphinxsearch.com/`

Table 2: Description of the Experiments at GikiCLEF 2009.

Automatic Runs	NLP in Query	Sphinx Match
gikiTALP1	-	MATCH_ALL (phrase rank)
gikiTALP2	lemma + stopwords filtering	MATCH_ALL (prhase rank)
gikiTALP3	lemma + stopwords filtering	MATCH_EXTENDED (BM25)

4 Results

The results of the gikiTALP system at the GikiCLEF 2009 Monolingual English and Spanish task are summarized in Table 3. This table has the following IR measures for each run: number of correct answers (*#Correct Answers*), *Precision*, and *Score*.

The run gikiTALP1 obtained the following scores for English, Spanish and Global: 0.6684, 0.0280, and 0.6964. Due to an unexpected error we did not produced answers for the Spanish topics in run 2 (gikiTALP2), then the results for English and global were 1,3559. The results of the scores of the run gikiTALP3 for English, Spanish and Global were 1.635, 0.2667, and 1.9018 respectively.

Table 3: TALP GikiTALP Results

run	**Measures**	**English (EN)**	**Spanish (ES)**	**Total**
run 1	#Answers	383	143	526
	#Correct answers	16	2	18
	Precision	0.0418	0.0140	0.0342
	Score	0.6684	0.0280	0.6964
run 2	#Answers	295	–	295
	#Correct answers	20	–	20
	Precision	0.0678	–	0.0678
	Score	1.3559	–	1.3559
run 3	#Answers	296	60	356
	#Correct answers	22	4	26
	Precision	0.0743	0.0667	0.0730
	Score	1.6351	0.2667	1.9018

5 Conclusions

This is our first approach for a Wikipedia-based retrieval task. We have used the Sphinx full-text search engine with limited Natural Language Processing processing and without using Geographical Knowledge. We obtained the best results when we have used all the NLP techniques (lemmas in the queries and stopwords filtered) and the Sphinx mode MATCH_EXTENDED. Geographical Knowledge as baseline algorithms. As a future work we plan to: 1) detect the Expected Answer Type and use the wordnet synsets to improve the results, 2) use Geographical Knowledge in the Topic Analysis, 3) increase the use of the Wikipedia links.

Acknowledgments

This work has been supported by the Spanish Research Dept. (TEXT-MESS, TIN2006-15265-C06-05). Daniel Ferrés is supported by a UPC-Recerca grant from Universitat Politècnica de Catalunya (UPC). TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

References

[1] Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, and Muntsa Padró. FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 48–55, 2006.

[2] T. Brants. TnT – A Statistical Part-Of-Speech Tagger. In *Proceedings of the 6th Applied NLP Conference (ANLP-2000)*, Seattle, WA, United States, 2000.

[3] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. Getting Geographical Answers from Wikipedia: the GikiP pilot at CLEF. In Francesca Borri and Alessandro Nardi and Carol Peters, editor, *Working notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September 2008. CLEF 2008 Organizing Committee.