Language Identification Strategies for Cross Language Information Retrieval

Alessio Bosca[1], Luca Dini[1]

[1] Celi s.r.l., Via S.Quintino 31
10131 Torino
{alessio.bosca, dini}@celi.it

Abstract. In our participation to the 2010 LogCLEF track we focused on the analysis of the European Library (TEL) logs and in particular we experimented with the identification of the natural language used in the queries. Language identification is in fact a key task within Cross Language Information Retrieval systems and the challenge is particularly difficult in the case of search queries where the contextual information available is scarce; function words (grammar particles highly connotative of a specific language like prepositions, pronouns, conjunctions, etc) are usually missing and the relevant presence of Named Entities can be misleading for the correct identification of the language used in the query. In order to face this challenge with acceptable performances the techniques applied should be different form the ones adopted for language guessing with more extensive and syntactically richer text fragments, like meta-data or textual documents. In particular we experimented combining together different strategies: corpus based, character model based and a priori hypothesis. Since no official evaluation of the task is available we manually evaluated a sample of 100 queries and the results obtained are quite promising.

Keywords: Cross-Language Information Retrieval, Language Identification, Log Analysis.

1 Introduction

The LogCLEF track proposes to investigate on the log analysis as a means to infer new knowledge and in particular the task proposes to the participants to deal with logs from The European Library (TEL). TEL provides access to national libraries of several European countries, therefore users and contents come from many languages and the logs provided in this task constitute a valuable opportunity and test-bed for evaluating Language Identifier strategies, specifically tailored to search queries.

In the last decade the demand for IT systems capable of integrating and correlating documents expressed in different languages generated a huge effort in the research community in order to support multilingual resources and Cross-Language Information Retrieval (CLIR) systems and different EU founded projects focused on this challenge, like Europeana[1], CACAO[2] or MICHAEL[3].

A key resource in order to support multilingual resources in IT systems is the capability of associating textual contents to the language in which they are expressed, whenever this information is not explicitly included within the meta-data associated to the resource itself. The same issue emerges in CLIR system supporting search queries translation as a mean to leverage multilinguality and provide access to all the documents satisfying user informational needs regardless of their language; the approach of querying in one language and retrieving documents in all the available languages is particularly significant whenever the contents of the exposed resources are not textual (images, audio, etc) and the constraint of being expressed in a specific language only concerns the meta-data.

Language Identification techniques traditionally (see [4], [5] or [6]) include models based on the statistical distribution of character sequences or the presence in the text of function words (grammar particles, highly connotative of a given language like conjunctions, pronouns, modifiers, etc) or comparing the frequency of terms in given language specific corpora. These different strategies present different needs with respect to available resources, computational power and processing time and yield different performances in different application context; therefore the most efficient approach in dealing with language identification would be selecting the technique with lower requirements for the given task.

Language Identification for search queries constitutes the most difficult task for language guesser components; in fact the contextual information available is scarce, function words (grammar particles, highly connotative of a specific language like prepositions, pronouns, conjunctions, etc) are usually missing and the relevant presence of Named Entities can be misleading for the correct identification of the language used in the query.

In our experiment we investigated the weighted combination of different strategies: corpus based, character model based and a priori hypotheses and applied these techniques to the user queries from TEL logs; since no official evaluation of the task is available we manually evaluated a sample of 100 queries and the results obtained are very promising.

This paper is organized as follows: we describe our experiments in Section 2 and present conclusions in Section 3.

2 Experiments Description

The first step in our investigations consisted in the extraction of all the distinct user queries from the TEL logs along with their ID and the associated UI language; this process resulted in about 450.000 user queries.

We then applied different language identification strategies to these list of search queries in order to evaluate their performances when applied singularly or combined together. Each software module implementing a specific language strategy returned as output a list of languages associated to a guess confidence value in the range of [0..1]. In particular we experimented the following strategies:

- *Pure Corpus Based*: languages are guessed comparing the frequencies of terms in the search queries within language specific corpora. The guess confidence value consists in the normalized sum of term frequencies.
- *Pure Character Model Based*: languages are evaluated comparing language model trained using textual contents from language specific corpora. The guess confidence represents the distance of the input text from a specific language model.
- The *Mixed Approach* combines together the two previous strategies with an even weight (0.5 Corpus Based, 0.5 Character Model Based)
- The *Mixed Approach with a Priori Hypothesis* introduces in the previous strategy a default guess, here represented by the UI language. In different application scenarios it could be the default language of the collection, the language retrieved from the user profile, etc. The weighting scheme used for this combined strategy is 0.4 Corpus Based, 0.4 Character Based, 0.2 A priori Hypothesis
- The *Mixed Approach without NE* investigates the effect on Language Identification performances of removing NE (when the search queries is not purely constituted of NE). Since a real NE recognizer module was unavailable for our experiments we emulated its presence exploiting the specific query syntax of TEL and removed the query terms pertaining to creator by means of the search field prefix “CREATOR ALL”.

Since no official evaluation of the task is available we manually evaluated a sample of 100 queries and the results obtained are presented in *Table 1*. All the search queries containing only Named Entities have been considered expressed in the language of origin of the referenced Named Entity (i.e. Hemingway ← 'en').
From the experimental evidence emerges that the most significant contribution to language identification came from the Corpus Based strategy although the contribution from the Character based approach can increase the overall performances.

Table 1. Language Identification Strategies.

Strategy	Correct Guesses	Wrong Guesses
Corpus Based	74	26
Character Model Based	57	63
Mixed Approach	76	24
with a Priori Hypothesis	75	25
without NE	4	2

Table 2 instead presents the statistical correlation of the language used in the search query and the language used in the User Interface; from the experimental evidence emerges the fact that the information on the language of the UI (here used as a priori hypothesis) is not more significant with languages different from the default one (here 'en'), therefore is not a relevant information to be used in order to increase performance of corpus based and character model based language guessing strategies.

Table 2. Query Language vs. UI Language.

UI Language	Same Language	Different Language
English	46	37
not English	8	9

3 Conclusion

The preliminary results are quite encouraging and in the future we plan to extend this research in order to include a full fledged Named Entity recognizer module.

Acknowledgments. This work has been supported and founded by *EuropeanaConnect* EU project.

References

1. Europeana project http://version1.europeana.eu/web/europeana-project/
2. CACAO project: http://www.cacaoproject.eu/
3. MICHAEL project: http://www.michael-culture.eu/
4. W. B. Cavnar and J. M. Trenkle, “N-Gram-Based Text Categorization”, In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval
5. B. Ahmed, S. Cha, "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", Proceedings of CSIS 2004, Pace University, May 7th, 2004
6. H. Ceylan, Y. Kim, "Language Identification of Search Engine Queries", Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1066–1074, Suntec, Singapore, 2-7 August 2009