Chapter XXXI
Retrieving Non–Latin Information in a Latin Web:
The Case of Greek

Fotis Lazarinis
University of Sunderland, UK

ABSTRACT

Over 60% of the online population are non-English speakers and it is probable the number of non-English speakers is growing faster than English speakers. Most search engines were originally engineered for English. They do not take full account of inflectional semantics nor, for example, diacritics or the use of capitals. The main conclusion from the literature is that searching using non-English and non-Latin based queries results in lower success and requires additional user effort so as to achieve acceptable recall and precision. In this chapter a Greek query log is morphologically and grammatically analyzed and a number of queries are submitted to search engines and their relevance is evaluated with the aid of real users. A Greek meta-searcher redirecting normalized queries to Google.gr is also presented and evaluated. An increase in relevance is reported when stopwords are eliminated and queries are normalized based on their morphology.

INTRODUCTION

According to recent statistics 64.2% of the online population, are non-English users (Global Internet Statistics, 2003). As the Web population continues to grow more non-English users will be amassed online. Recent studies showed that non-English queries and unclassifiable queries have nearly tripled since 1997 (Spink et al., 2002). Even though several Web search engines exist, most of their features and virtues are catered for the English language only. For example, the query “Bookshop New York” in Google retrieves Web pages mentioning the semantically related words “book”, “books” and “bookstore” as well. This is easily understood as the matching terms are emboldened. In contrast, the queries “Librairie Paris” in French, “Libreria Roma” in Italian, “Librería Madrid” in Spanish and
“Βιβλιοπωλείο Αθήνα” in Greek, retrieve only pages which include exactly the query terms as they are typed in the query. Another more convincing example results from the query “stemming site:www.dcs.gla.ac.uk/Keith” in Google. Here the search is restricted to Keith’s van Rijsbergen site. Three pages are returned and both the words “stem” and “stemming” are emboldened as they are considered matching terms. The Greek query “Βιβλιοπωλείο Αθήνα” (Bookshop Athens) retrieves 37,300 pages in Yahoo, while its upper case equivalent “ΒΙΒΛΙΟΠΩΛΕΙΟ ΑΘΗΝΑ” retrieves 2,760 pages with low correlation among the top ranked results between the two queries versions. In English Web searching the upper and lower case version of the same query are handled consistently in all search engines. This is true for other Latin languages as well. Additionally, Latin queries where diacritics are omitted are handled effectively in major search engines like Google and Yahoo. For instance, the queries “Università Roma” and “Universita Roma” retrieve exactly the same number of pages in Google ranked identically. Yahoo acts correspondingly in these two Italian queries as well. In contrast, the Russian words “рабочей” (worker) and “рабочеи” retrieve respectively 7,020,000 and 1,680 pages in Google. The Serbian queries “Социјалистичка Федеративна” (Socialistic Federation) and “Социјалистичка Федеративна” retrieve 682 and 811 Web pages respectively in Yahoo. The results indicate that often the accent marks are omitted and therefore search engines should consider this fact and compensate for omitting of diacritics.

The problems mentioned previously affect the text based image retrieval facilities of the search engines and probably all of their services. In Google the Russian words “рабочей” (worker) and “рабочеи” retrieve 195,000 and 23 images respectively. The Greek queries “ΣΚΥΛΟΣ” (dog) and “σκύλος” retrieve 210 and 1000 images respectively. All the queries presented up to now were run on the same day in April 2007 to ensure compatibility of the results and validity of the inferences.

The aforementioned examples show that the international search engines manipulate uniformly the upper and lower case English and Latin queries and the absence of the diacritics. They also propose semantically related terms to augment the retrieved set of pages and to improve ranking and employ information retrieval techniques such as stemming (Baeza-Yates & Ribeiro-Neto, 1999). More options such as query suggestions based on misspelled user requests and efficient handling of stopwords (Baeza-Yates & Ribeiro-Neto, 1999) are available to English users. Also services like Google’s Scholar [http://scholar.google.com] or Yahoo video search [http://video.yahoo.com] operate practically only on Latin named resources and Web pages.

To effectively support the information needs of non English and non Latin Web searchers, we need primarily to understand how users interact with search engines and to thoroughly study their queries. Then the relevance of queries following specific patterns should be evaluated. Finally, in order to improve Web searching in a specific natural language new tools and techniques should be proposed taking into account the linguistic features and restrictions of this language.

**NON-LATIN WEB SEARCHING**

Non-Latin Web queries are continuously increasing as the Internet penetrates Asian, African and eastern European countries which do not use the Latin alphabet. Existing activities like CLEF [Cross-Language Evaluation Forum - http://www.clef-campaign.org] and NTCIR [NII Test Collection for IR Systems - http://research.nii.ac.jp/ntcir/] focus on core information retrieval techniques for improving retrieval of non English queries in closed corpuses but they are not sufficiently focused on the requirement to build better search engines for all forms of non-English queries.
Related Content

A Hyper-Heuristic for Descriptive Rule Induction
[www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a](www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a)

An HL7-Aware Decision Support System for E-Health
[www.igi-global.com/chapter/hl7-aware-decision-support-system/21750?camid=4v1a](www.igi-global.com/chapter/hl7-aware-decision-support-system/21750?camid=4v1a)

Cost Models for Selecting Materialized Views in Public Clouds
[www.igi-global.com/article/cost-models-for-selecting-materialized-views-in-public-clouds/117156?camid=4v1a](www.igi-global.com/article/cost-models-for-selecting-materialized-views-in-public-clouds/117156?camid=4v1a)

User Behaviour Pattern Mining from Weblog
[www.igi-global.com/article/user-behaviour-pattern-mining-weblog/65571?camid=4v1a](www.igi-global.com/article/user-behaviour-pattern-mining-weblog/65571?camid=4v1a)