Improving Non-English Web Searching

(iNEWS07)

Fotis Lazarinis
Technological Educational Institute
Mesolonghi, Greece
lazarinf@teimes.gr

Jesus Vilares Ferro
University of A Coruña
Spain
jvilares@udc.es

John Tait
Information Retrieval Facility
Austria
john.tait@ir-facility.org

Workshop website: http://rea.teimes.gr/~lazarinf/ir7w/

Abstract

This workshop attempted to promote the discussion and the research on non-English Web searching. Most search engines were first built for English. They do not take full account of inflectional semantics nor, for example, diacritics or the use of capitals. Our main aim was to discuss the additional problems faced in non-English Web queries and to suggest techniques to improve the response of searching systems. Papers related to Arabic, Basque, Farsi (Persian), Greek, Spanish, Swedish, Hindi, Bengali and other south asian languages were accepted. Conclusions were that search engines would be more effective if they took more account of the properties of individual languages, and that there is a need for more studies of real user behaviour in practical situations.

1 Introduction

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. Recent studies showed that non-English queries and unclassifiable queries have nearly tripled since 1997. Most search engines were originally engineered for English. As a result they do not take full account of significant features of languages which are absent or unimportant in English. Such features include inflectional semantics diacritics or the use of capitals in individual languages.

The main conclusion from the literature is that searching using non-English and non-Latin script queries results in lower success and requires additional user effort so as to achieve acceptable recall and precision. Further international search engines (like Yahoo and Google) are relatively weaker with monolingual non-English queries.

The main goals of the "Improving Non-English Web Searching" (iNEWS07) SIGIR 2007 workshop were to evaluate search engines in non-English queries; measure the additional user effort; and propose solutions to the identified problems.

2 The Workshop

Around 25 researchers attended the one day iNEWS07 workshop. The researchers came from various academic institutes from Germany, Greece, Iran, Spain, Sweden, UK, USA and from industry (e.g. Microsoft).

In response to our call, after a triple blind reviewing process, 4 papers were selected as full papers (references 1-4), and 6 more as short papers/posters (references 5-10) covering different languages like Arabic, Basque, Farsi (Persian), Greek, Spanish, Swedish and a variety of Indian and South Asian languages which use the Abugida script systems, including Hindi and Bengali.

The day began with a short presentation of the aims of the Workshop and background information on non-English Web searching. Next, Professor Maarten de Rijke (University of Amsterdam) presented his invited talk focused on the monolingual non-English queries at the WebCLEF track and the research at the University of Amsterdam. Full papers were presented in the subsequent part of the event. The next session was dedicated to short papers. The last part of the workshop was a discussion session to identify the common problems related to Web searching across different non-English languages.

Presenters of full papers were given 30 minutes to present their papers and to discuss their research. Authors of short papers had to prepare a poster and were asked also to make a short two-slide presentation with the main points of their work followed by a discussion.

2.1 Session I: Invited talk

The workshop opened with an invited talk "*Who's the user? Who's the researcher?*" by Professor Maarten de Rijke from the University of Amsterdam. In his talk Professor de Rijke mentioned that "Over the past few years there has been a lot of progress in technology used for addressing monolingual or multilingual web queries in languages other than English. Nevertheless, a great deal of work still remains to be done, e.g., on the morphological analysis of non-English web queries, before the retrieval performance on English and non-English are on a par. There is another pressing issue, however, that is at least as important: we know very little about users of monolingual or multilingual (non-English) web search facilities. Who are they? What do they search for? What are their intents? At WebCLEF (http://www.clef-campaign.org/) - the multlingual web retrieval track run at CLEF - these questions and concerns have led to a very explicit definition of the retrieval task, where various assumption are being recorded as part of the topic statement."

In his talk Professor de Rijke reviewed the choices made at WebCLEF over the past few years and detailed the current set-up. Another important aspect of the talk concerned the issue of lack of user data that most academic research groups have to work with. He discussed various ways around this, one example being the use of publicly available and usable showcases and demonstrators. University of Amsterdam runs a small number of Dutch language online search and browsing tools. A number of findings of this strategy were also presented, based on a brief log analysis together with both quantitative and qualitative analyses.

2.2 Session II: Full Papers

Four papers were presented in the second session concerning Greek Web searching, non-English queries in Terrier and a model for automatic language identification.

The first paper [1], presented by Professor Efthimiadis, focused on evaluating how well search engines respond to Greek queries and on how to compare and assess the effectiveness of search engines in Greek queries. The presenter explained the characteristics of Greek and the morphology of Greek text. It was also explained that the terms transcription, transliteration and romanization refer to the mapping of non-Latin text to Latin-based text. In their study, Efthimiadis et al. [1], run a number of Greek and transliterated queries in Greek and in international search engines and checked the top ten results against predefined lists of relevant websites. Their intention was to test whether the search engines were able to discover the websites which are known in advance to be relevant. Their results show that although global search engines are faster and with richer indexes they treat differently Greek queries of equivalent meaning but with different morphology. On average the search engines retrieved the desired document in the first three rank positions which leaves much space for improvement.

Hammarström presented a model to identify the language of a written document given a set of candidate languages and training data for them [2]. The traditional approaches based on n-gram frequency tables were initially discussed. The proposed model combines frequency dictionary and unsupervised affix detection. These two components are combined into a word emission probability distribution that aims to predict how likely a language is to have emitted a given word. The evaluation showed that the new proposed model is bigger and slower in practice than earlier methods but it could benefit a special class of practical applications such as spell checkers and can work with many natural languages.

The performance of Terrier was the focus of the next paper [3] presented by Craig Macdonald. The aim of this work was to identify how standard Information Retrieval techniques can be adapted in Web retrieval for non-English queries. The challenge of stemming queries and documents in a multilingual setting was addressed. Experiments with a multilingual collection of over 20 languages, more than 800 queries, and various stemming strategies in these languages revealed that using no stemming results in satisfactory Web retrieval performance. Moreover, it was shown that language specific stemming requires an accurate identification of the language of each query.

The last full paper [4] presented by Paraskevi Tzekou focused again on the Greek language and more specifically on Greeklish queries. Greeklish queries are romanized (transliterated) text. Although Greeklish emerged as a convenient mean for the creation and distribution of digital data at a time when Unicode Transformation Format (UTF) was not supported for the Greek alphabet, nevertheless it is still being utilized as a matter of habit or need. In their paper, Tzekou et al. [4] proposed a model that treats Greek and Greeklish web data in a uniform manner. Their aim was to improve the usability of Greek search engines and ameliorate the user experience, regardless of the preferred query alphabet. It was shown that about 46% of the participants of an evaluation study, issue Greeklish queries when looking for web resources. It was further suggested that 40.5% of the subjects use Greeklish queries when their Greek searches fail to retrieve the desired information. The paper proposed the conflation of Greek and Greeklish data in the searches and their experiments proved that expanding Greeklish query terms with their Greek equivalents increases the relevance of the search results.

2.3 Session III: Short Papers/Poster

Six papers were selected as short papers/posters. Apart from the poster a two-slide presentation and a discussion was required for each short paper.

An n-gram conflation approach for Arabic text was the topic of paper [5] presented by Farag Ahmed. In this paper a language independent approach for conflation that does not depend on predefined rules or prior knowledge in the target language was presented. It was shown that the proposed method is effective to achieve high score similarities between all of the word form variations. It also reduces the ambiguity and obtains a higher precision and recall, compared to the pure n-gram based approaches.

Igor Leturia presented EusBila [6], a search service for Basque that relies on the APIs of search engines, yet obtains a lemma-based and language-filtered search by means of morphological query expansion and language-filtering words. The authors argue that using standard search engines for making a query in a minority and agglutinative language like Basque is unsatisfactory in terms of precision. EusBila uses the indexes of other search engines and limits the results in Basque by using language-filtering words.

The next paper [7] was presented by Ernesto William De Luca and concerned multilingual query-reformulation. Their tool supports the learner/user to find all possible word senses and then retrieve and categorize documents. Eventually users formulate multilingual queries, giving them the possibility to explore the intended meanings in other languages. The word senses of different word combinations can be disambiguated by translating them using EuroWordNet.

Behrang Qasemizadeh presented a paper about e-orthography in Farsi (Persian) [8]. Farsi uses a unified Arabic script as its writing system and there is no standard format for Farsi orthography in the digital environment. Therefore the author proposed an approach to represent Farsi electronic texts called e-orthography. E-orthography indicates how the orthography of a language can be followed within an encoding system. The author argues that including the e-orthography concept as a part of search engines design can enhance recall and precision.

The next sort paper [9] presented by Víctor Darriba proposes a new method for assigning topics from a hierarchical thesaurus to documents written in natural languages. Their approach models thesaurus topic assignment as a multiple label classification problem, where the whole set of possible classes is hierarchically organized. The developed system has a high performance on average documents but in complex or very specific documents it is unable to reach the expertise of a human.

The authors of the last paper [10] unfortunately did not attend the workshop. This paper concerned text searching for languages using Abugida scripts.

2.4 Session IV: Discussion Session

The keynote and the papers presented in the previous sections revealed several of the problems which influence the precision and increase the user effort in non-English Web searching. The last session, chaired by Professor John Tait, aimed at identifying the most important Web searching problems which are common across various natural languages.

The participants agreed that the performance of search engines is not as high as it is expected and most of the global search engines ignore the morphology of the queries in many natural languages. Therefore users should be more knowledgeable to increase the possibility of retrieving more relevant results. It is not certain that techniques such as stemming would help and therefore more sophisticated techniques such as lemmatization and query reformulation are needed to increase the precision of search engines. Problems also arise from the encoding of text, especially in non-Latin script languages. Further, in many websites text is transliterated and some users use Romanized versions of Web queries. This user behaviour should be taken into account by search engines in order to help users.

3 Conclusions

The main conclusion from the workshop is that there are still a lot of improvements needed in order to provide more accurate searching to non-English users. Search engines should become more aware of the individualities of each natural language. User behaviour should be further studied to understand how they form their queries. Finally work is needed for studying what techniques could help them improve the precision of their results.

4 References

1. Efthimiadis E., N. Malevris, A. Kousaridas, A. Lepeniotou and N. Loutas (2007), How do Search Engines handle Greek Queries? In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 9-13.
2. Hammarström H. (2007), A Fine-Grained Model for Language Identification In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 14-20.
3. Macdonald C., C. Lioma and I. Ounis (2007), Terrier takes on the non-English Web, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 21-28.
4. Tzekou P., S. Stamou, N. Zotos, E. Kozanidis (2007), Querying the Greek Web in Greeklish, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 29-38.
5. Ahmed F., A. Nürnberger (2007), N-Grams Conflation Approach for Arabic Text, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 39-46.
6. Leturia I., A. Gurrutxaga, N. Areta, I. Alegria, A. Ezeiza (2007), EusBila, a search service designed for the agglutinative nature of Basque, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 47-54.
7. De Luca E. W., M. Eul, A. Nürnberger (2007), Multilingual Query-Reformulation using RDF-OWL EuroWordNet, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 55-61.
8. Qasemizadeh B. (2007), Farsi e-Orthography: an Example of e-Orthography Concept, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 62-64.
9. Ribadas F., E. Lloves-Calvino, V. M. Darriba (2007), Thesaurus topic assignment using hierarchical text categorization In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 65-70.
10. Singh A. K., H. Surana, K. Gali (2007), More Accurate Fuzzy Text Search for Languages Using Abugida Scripts, In: F. Lazarinis, J. Vilares, J. Tait (eds) *Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop*, pp. 71-78.