Across the Bridge: CLEF 2001 – Non-English Monolingual Retrieval. The French task.

Eugenia Matoyo & Tony Valsamidis,
University of Greenwich,
Old Royal Naval College, Park Row, Greenwich, London SE10 9LS
me010@gre.ac.uk or a.Valsamidis@gre.ac.uk

This paper presents work on document retrieval based on participation in the Cross-Language Evaluation Forum (CLEF) 2001 task of non-English monolingual retrieval task using French only. In summary, the experiment findings indicate that Okapi, the text retrieval system in use, can successfully be used for non-English text retrieval although a lot of internal pre-processing is required in the basic search system to convert the documents and topics into Okapi access formats. Various shell scripts were written to achieve the conversion in a Unix environment, failure of which would significantly have impeded the overall performance. Based on the experiment findings using Okapi, which was originally designed for the English Language, it was clear that there was significant difference between French and English retrieval depending on the retrieval system in use.

INTRODUCTION

Participation in the CLEF 2001 monolingual task [2] for French serves as a stepping stone toward the multilingual cross-language information retrieval. The background to the CLEF Campaigns as presented in the European Research Letter (2000)[4] triggered interest toward participation in this year's information retrieval task, which was our first attempt to participate in any of the CLEF tasks. The main task attempted was that of monolingual (non-English) Information Retrieval on French topics and documents as assigned by CLEF and retrieval was based on automatic query construction using Okapi. The main reason for choice of automatic query retrieval as opposed to manual query expansion was largely the time constraint. Automation implies that documents were automatically retrieved by the system without the intervention of the experimenter in interactive manual query expansion[9], which would have been time consuming and time was a limited resource given that the experiments had to be completed in time for the CLEF submission deadline.

Objectives

The primary aim of this experiment was three-fold:

> Crossing the Bridge: An attempt at participation in the CLEF tasks on Information Retrieval for better understanding of issues in Cross-language information retrieval.

To investigate whether Okapi, an experimental text retrieval system from City University of London[1], could successfully be used to provide a useful interface in the retrieval of French documents using indexing methods and stopwords approach.

To investigate whether techniques applied for English text retrieval differ significantly from those used for French retrieval.

GENERAL SYSTEM DESCRIPTION

The monolingual experiments for French documents were carried out using Okapi, a text retrieval system project based at City University, London, which is used solely for research purposes. The Okapi system requires either a Solaris on Sun environment or Linux on Intel environment to run the software. It uses a probabilistic model[10] of information retrieval, which was first developed by Robertson[7]. This model performs in an iterative process, which uses the ranking of document listings based on indexing, term weighting function and word stemming rules for optimizing search queries. The three major components of Okapi are:

Indexing Software, which enables users to create and index Okapi type databases

The Basic Search System (BSS), which is a set of low level commands to enable users to build their own interface around it.

Okapi Interactive Interface, which is the graphical user interface, which calls BSS command in a manner such as to hide the complexity from the user.

It is important to point out here that Okapi system was originally designed for use with English. Although it had been used in similar text retrieval experiments for languages other than English[9], using it for the CLEF experiment for monolingual tasks was the very first time that it has been used for French. The section below describes the work that had to be done in the Okapi basic search system to allow for French monolingual retrieval.

PRE-PROCESSING TECHNIQUES

Prior to carrying out the formal runs, it was necessary to carry out various pre-processing tasks on the topics and the French collection using various shell scripts based on previous experiments in the Text Retrieval Conference (TREC)[6] series.

Database Integration

The two separate French collections, *le_monde.tar.gz* (Initial zipped file of 154 MB) and *sda_french* - Initial zipped file of 80 MB), were integrated into a single database, *le_m+dsa_fr* upon which to carry out the query processing.

Conversion

A shell script (convert_topic) was written to reformat the 50 CLEF French topics by altering the title from *Fr-Title* to *Title* (because retrieval was required for the T*itle* field) and changing the document numbers by removing the preceding *CO*. This reconstruction and manipulation of the topics and documents was necessary to convert the topics into Okapi access format as well as to enable the reuse of previous shell scripts, which had been used in similar TREC experiments. Various changes to the scripts were, however, required to customize them for the CLEF collection for French. Below is the script *convert_topic:*

```
# Script to convert CLEF French topics to TREC  format
sed -e 's/FR-//g' \
   -e 's/C0//'
#Substitute nothing for FR globally from the topic titles to be plain <title>
# Substitute nothing for CO globally for the Document numbers to be in plain numbers.
```

Stoplist

The *stoplist*, also referred to as *stopwords* or *stopterms*, is made up of common, generally used words in a language collection, which are considered irrelevant for the purpose of information retrieval because of their high frequency of occurrence. The stoplist comprised 248 terms and was derived both from several sources to include common French terms[3], information compiled by Professor Jaques Savoy of the University of Neuchatel for the CLEF web site[5] and based on Prof. Savoy's publication in the Journal of the American Society for Information Science[8]. All accented words were removed from the stoplist because Okapi cannot handle accented words.

A shell conversion script was required to convert the stoplist to Okapi-accessible formats because a *H* was required before each word, a comma [,] after each word and finally a colon [:] at the end of each line. An example of the conversion is from *avec* to *Havec,:* to comply with Okapi formats. This was achieved using an emacs editing utility in the Unix environment to write a simple conversion script:

```
#For every line
#French stoplist
Replace string ^J with : H  # H to indicate stop term
Replace string ^J with : ,:  # to conform with Okapi acceptable for stopterms
```

Stemming

There was no stemming applied at all for this experiment. Porter's stemming algorithm, which is configured to work for Muscat - an open source search engine - had been intended for use with Okapi but failed to run successfully. The Muscat stemmer depends on its own character coding, which Okapi could not recognize because it does not work with unicode. A *frontend* shell script was written to adapt Okapi Basic Search System (BSS) to the stemmer but this also failed to work. Given more time, the script may have been successfully debugged and configured for Okapi but this was not accomplished due to time limitations. Since Porter's stemmer for English has successfully been used previously with Okapi in similar TREC experiments[9], it implies that Okapi could not properly handle the accented words in the French language for the stemmer to work in this experiment. Thus stemming was omitted altogether and it may be worthwhile to note that this lack of stemming has resulted in less accurate results in the final formal runs for the experiment.

Indexing

Indexing of terms was accomplished in Okapi using an inbuilt Okapi indexing utility to index the integrated collection. The utility was called using two of Okapi's BSS (basic search system) programmes *ix1* and *ixf* as shown below:

ix1 -delfinal le_m+sda_fr 1 | ixf le_m+sda_fr 1

The same programmes were used to index the 50 CLEF topics for the combined collection:

ix1 -mem 50 -delfinal -doclens le_m+sda_fr 0 | ixf le_m+sda_fr 0

Testing - Comparison with CLEF input checker

Throughout the experiments, it was important to ensure that the results would be in the required format for CLEF. The test runs were, therefore, validated against the CLEF input checker, CheckInput.pl. This was a Perl script, which compared the run results against the input checker for error identification in an effort to expose possible defects in the runs before submission of the formal runs. The input checker revealed that there were no serious errors in the run results, although a couple of topic sets yielded error on retrieval, possibly as a result of a segmentation bug in the Okapi BSS release.

THE FORMAL RUNS

The formal runs were automatic and query processing for the French topics and documents was done in three separate runs, each using the the same French stoplist:

Run 1 - retrieval of documents by topic Title, Description and Narrative

Run 2 - retrieval by topic Title and Description only

Run 3 - retrieval by topic Title only

The graph below indicates the order of importance as assigned by the system, and retrieval by Title, Description and Narrative received higher ranking:

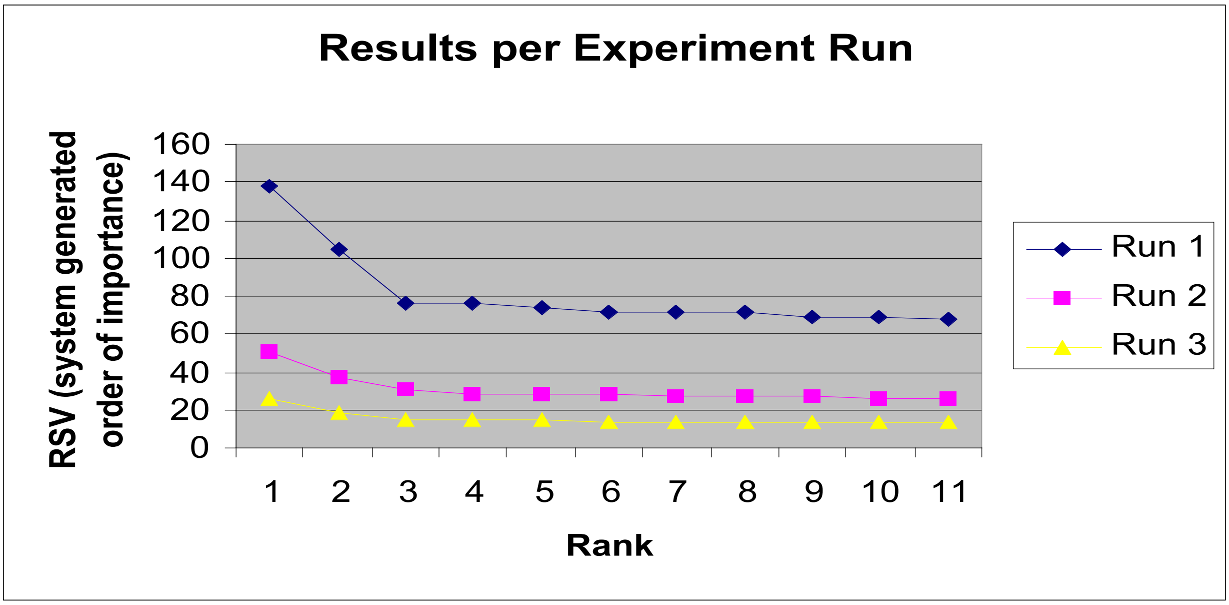

Fig 1.0 Chart showing Results per experiment run for the same topic and ranked documents

Where: *Run 1 represents run gre1 for retrieval by Title, Description and Narrative*

Run 2 represents run gre2 for retrieval by Title and Description

Run 3 represents run gre3 for retrieval by Title only:

As expected and observed from the Fig 1.0 respectively, all the three runs differed significantly from each other. Retrieval by *Title*, *Description* and *Narrative* gave the highest document relevance score of the three runs and retrieval by Title only of the same documents yielded the lowest document relevance score.

PRESENTATION AND RESULTS

The formal runs were zipped using gZip in Unix and were submitted by FTP to CLEF in ASCII as binary format and the results were sent out together with a README file describing the different runs. The formal results from CLEF for the submitted runs were as expected; Run *gre3*, which had the least priority was not judged. Similarly, in a graph of recall versus precision, run *gre3* scored the least precision values. Our results for the experiment can be summarised as:

```
CLEF evaluation results for run gre1:
=====================================================================

This run was JUDGED, i.e. the results file contributed to the
relevance assessment pool.

CLEF evaluation results for run gre2:
=====================================================================

This run was JUDGED, i.e. the results file contributed to the
relevance assessment pool.

CLEF evaluation results for run gre3:
=====================================================================

This run was NOT JUDGED, i.e. because of limited evaluation resources,
the result file did not directly contribute to the relevance assessment
pool. However, the run was subject to all other standard processing,
and is still scored as an official submission.
```

Fig 2.0 Summary of Run Results

Although all of our experiment runs had a relatively poor performance against the comparison to the 'median' graph by topic from CLEF, which gave an indication of how well our results were according to other groups, run *gre1* had a considerably better performance than both runs *gre2* and *gre3*. Run *gre3* had the worst median performance of the three runs. The average precision (non - interpolated) for all relevant documents was 1.0000 for both runs *gre1* and *gre2* but only 0.3381 for run *gre3*. Our deduction from the runs thus shows that document retrieval using the topic *Title*, *Description* and *Narrative* fields yields far better results than attempted document retrieval by omission of any of these fields.

CONCLUSIONS AND FUTURE WORK

Work done on similar TREC experiments proved to be useful for the CLEF experiments by enabling reuse of previous scripts. Failure of Okapi to properly handle accents in the French language leads to our conclusion that procedures for monolingual information retrieval are not completely language independent. Different languages present different problems depending on the information retrieval system in use. Working with a French collection does not guarantee the use of established methods (such as indexing and stoplists), which would work well with English text retrieval. Methods that may be highly efficient for certain language typologies may not be so effective for others. Future work would involve adapting Porter's stemmer to Okapi and attempting the cross-language information retrieval in the multi-lingual tasks. We did not really expect great results, this being the first time round, but it is our hope that future participation will considerably yield better results, especially after sharing experience with other participants at the CLEF Workshop.

Acknowledgements

Without *Steve Walker*, the inventor of Okapi, we would not have been able to meet the CLEF deadline for participation in the 2001. He carried out a large proportion of the pre-processing activities for data conversion into access formats for both Okapi and CLEF. He also handled the testing of our formal runs against the CLEF input checker based upon his previous experience from participation in the Text Retrieval Conference (TREC)[6] series.

Special acknowledgement of help and collaboration also goes to *Surinder Singh Dio* of the Unix help at the University of Greenwich. He set up the okapi pack in the computing laboratory at the university and took great pains to set up the environment variable settings as required for the successful operation of the Okapi software.

References:

1. Centre For Interactive Systems Research, City University: *Introduction to Okapi Pack*, 14 March 2000, available from: http://dotty.is.city.ac.uk/okapi-pack/okapi-pack.html (accessed 17 April 2001)
2. Cross Language Evaluation Forum, CLEF: *CLEF Agenda for 2001*, 14 March 2000, available from: http://www.iei.pi.cnr.it/DELOS/CLEF/clef01.html (accessed May 2001)
3. Mueller, Erik T. 1998. *Fluent French: Experiences of an English speaker*. New York: Signiform. Available: http://www.signiform.com/french/ (accessed June 4, 2001)
4. Peters, C. & Braschler, M. (forthcoming). *Cross-Language System Evaluation: the CLEF Campaigns.* European Research Letter. To appear in Journal of American Society for Information Science and Technology.
5. Savoy, J. 2001, Stopword List, available from : http://www.unine.ch/info/clef/ (accessed June 5, 2001)
6. Text Retrieval Conference, *Test Collections*, available from: http://trec.nist.gov/ (accessed May 2001)
7. Centre for Interactive Systems Research, *The Probabilistic Retrieval Model*, available from: http://www.soi.city.ac.uk/research/cisr/okapi/prm.html (accessed May 2001)
8. Savoy J. 1999. *A stemming procedure and stopword list for general French corpora.* Journal of the American Society for Information Science, 50(10), 944-952.
9. M.Beaulieu, *Experiments on Interfaces to Support Query Expansion* (p8-19). S.E. Robertson, S. Walker and M. Beaulieu, *Laboratory experiments with Okapi: participation in the TREC programme* (p20-34). S.E. Robertson and M. Beaulieu, *Research and evaluation in information retrieval* (p51-57). X. Huang and S.E. Robertson, *Application of probabilistic methods to Chinese text retrieval* (p74-79). Special issue of Journal of Documentation 53(1), (1997).

10. K. Sparck Jones, S. Walker and S.E. Robertson, *A probabilistic model of information retrieval: development and status*. Available from: http://citeseer.nj.nec.com/jones98probabilistic.html