Improving Cross-lingual Search Quality

E. Sizikova, I. Ibarra, C. Quaranta, E. Schwartz, R. Bandari,
Research in Industrial Projects for Students (RIPS). University of California, Los Angeles (UCLA)

Abstract

Advancements in Information Retrieval (IR), a field that has become drastically more important in the Information Age, focus primarily on increasing the speed and accuracy of search upon large collections of data. We present several methods for cross-lingual search in the Shoah Foundation Institute Visual History Archive. The nature of our project yielded itself to value the recall performance higher than the precision performance of a given method. Therefore, the majority of the developed techniques focus on obtaining relevant terms from a variety of sources on the web, and expanding our database of results.

Sponsor: The USC Shoah Foundation Institute

- Established after *Schindler's List*.
- Stephen Spielberg was getting lots of phone calls.
- Over 52,000 interviews of survivors and witnesses of the Holocaust.
- Indexed by approximately 62,000 English keywords.
- Worldwide access for education.

Baseline Method

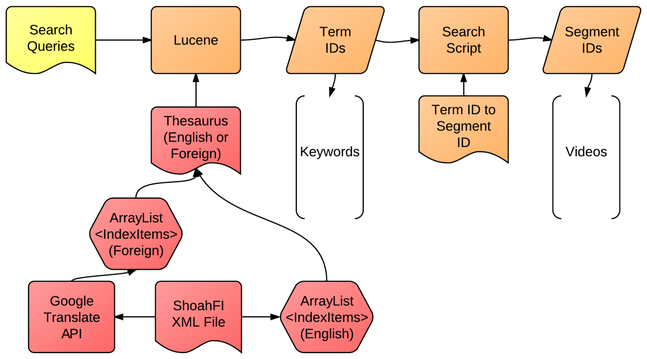


Figure: Our initial idea was translate the labels, and translate the search keyword, then use language specific implementation of Lucene to search.

Baseline Method Performance

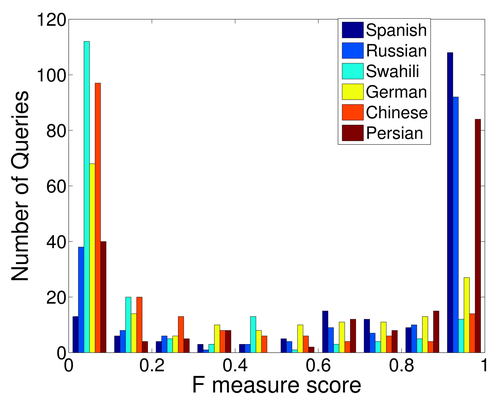


Figure: Baseline method had a clear division between languages that returned successful results majority of the time, and languages that did not.

Known problems:

1. Some languages had better stemming and analyzing support than others
2. 'Double translation' through Google leads to more problem queries
3. Language-specific implementation problems
4. Difficult to set-up for a website that requires quick results

SQTM Method

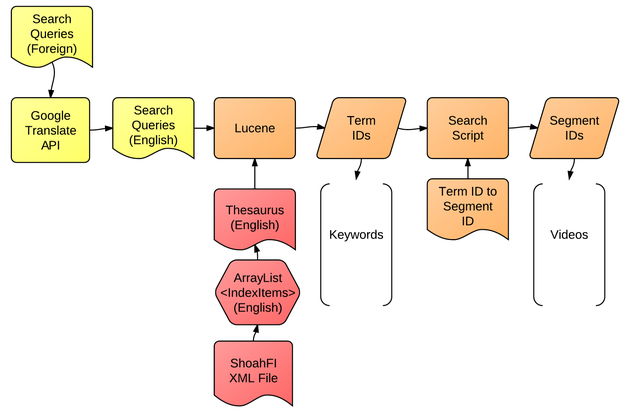


Figure: The next method was to directly translate the search query, and use English Lucene to search in the database.

Benefits of using SQTM:

1. Only one machine translation is performed: this amounts to less possibility for error.
2. SQTM more easily lends itself to various typical techniques that can be used to improve our results.
3. Every language is passed through the same Lucene version (only English), which provides an easier way to implement the system.
4. The number of languages is limited by the number of languages supported by Google translate or similar software, we are no longer concerned how many languages Lucene stemmers and analyzers can work with.

Possible Improvements

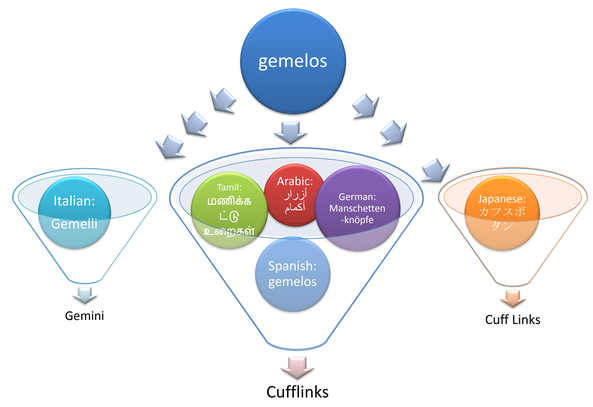


Figure: Improving SQTM results is difficult since neither precision not recall is perfect. We need a method that would not increase recall at a big expense of precision.

Improving Precision: Filter

Figure: Irrelevant words are filtered out before the search.

Improving Recall: WordNet

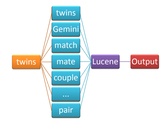
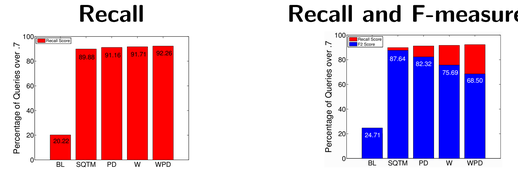
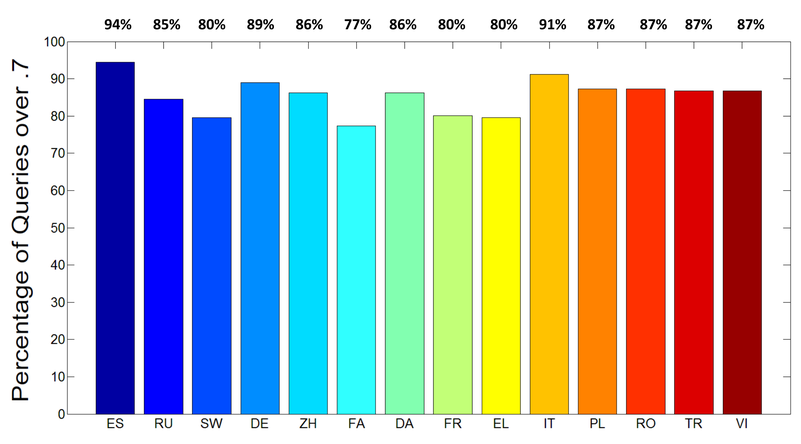


Figure: Each query is first sent to WordNet and then each returned synonym is fed into Lucene to be searched.

Comparison of Methods: Improved Recall and Declining Precision



SQTM: All Languages



Conclusions and Further Work

Our results clearly indicate that SQTM is the best method for solving the specific problem posed by the Shoah Foundation. To see the full stength of pivoting, filtering, and using other corpora, we need to obtain a more complete thesaurus of search terms. For example, one possible improvement to the thesaurus and hence to the search quality to use a wider range of transliterations of words.

Acknowledgements

This project was jointly supported by the USC Shoah Foundation and NSF Grant # 0931852. We would like to thank Prof. Mike Raugh and Prof. Russel Caflisch for their support at IPAM, as well as Sam Gustman, Leo Hsu, and other staff for their support at the Shoah Foundation. Finally, thanks to Women in Machine Learning (WiML) and their sponsors for sponsoring this poster.