iCLEF at Sheffield

Zoë Bathie, Mark Sanderson (m.sanderson@shef.ac.uk)
Department of Information Studies, University of Sheffield,
Western Bank, Sheffield, S10 2TN, UK

Abstract

Sheffield's contribution to the interactive cross language information retrieval track took the approach of comparing user's abilities at judging the relevance of machine translated French documents against ones written in the users' native language: English. Conducting such an experiment is challenging, and the issues surrounding the experimental design are discussed. Experimental results strongly suggest that users are just as capable of judging relevance on the native language documents are they are on the translated.

Introduction

An important and relatively little studied aspect of cross language information retrieval (CLIR) research is user interaction with such a system. Even the most fundamental aspects of retrieval, such as user ability to formulate effective queries or judge retrieved documents, has hardly been examined in a cross language context. As a consequence, the interactive cross language evaluation forum (iCLEF) was set up. Starting this year, the track studied one aspect of the interactive process: the user's ability to judge the relevance of retrieved foreign (i.e. target) language documents translated in some manner into the users' native (i.e. source) language. Using a test collection, user relevance judgements were compared to the judgements previously made by relevance assessors. The aim of the track was to compare different translation methods. However, at Sheffield, a different approach was taken: comparing user ability to judge the relevance of translated news articles against ability to judge articles written in the user's native language. The rest of this article describes the Sheffield work: a short literature review prefixes the experimental design followed by a discussion of issues arising from the design. Next, the results are presented, and possible future work is outlined.

The experiment

It might seem reasonable to assume that if a user is presented with a clearly written document that another has judged for relevance, the user will agree with the judgement. However, as is well known, relevance assessments are subjective depending on user interpretation of the query and document, which is based on prior knowledge of the subject. Consequently, there can be a reasonable level of disagreement between judges. Voorhees (1988), amongst others, studied this issue.

The iCLEF experiments compare relevance judgements of users against those made previously by assessors, although in this case the assessors were reading the judged documents in their original language and the versions examined by the users were translations of some type. Although the aim of iCLEF was to assess the extent the translation had impaired users' ability to judge, any such measurement would also include disagreements between users and assessors on what constitutes relevance. Others conducting the iCLEF experiment choose to rely on past work on levels of disagreement to provide an indication of how important this factor is. Sheffield opted instead to attempt to separate out these factors by conducting a form of control experiment: comparing ability to judge relevance of translated documents against judgements made on native language documents. Specifically, users were presented with documents retrieved in response to iCLEF test collection queries: from either French newspaper articles automatically translated using Systran software; or English language articles from the LA Times. Both newspapers collections covered the same time frame. Specifics about the queries used, collections searched and forms of relevance judgement made are outlined in the iCLEF overview paper elsewhere in the iCLEF notes (Oard & Gonzalo Arroyo 2001).

Designing an experiment to compare effectively user ability to judge relevance in native and translated documents is problematic. In the design chosen here any difference in relevance judgements across the two sets documents can be attributed to factors other than the quality of English in the texts. Writing styles or assumptions of prior cultural knowledge may differ in Le Monde and the LA Times and such factors may affect user relevance judgement. In addition, the assessors (to whom user judgements are compared) are different for the two collections, as are the conditions under which they performed their assessment; again this might be an influencing factor. Even the retrieval system may have behaved differently on the two collections and this may influence the type of retrieved relevant documents presented to the user. Despite these issues, it was judged that continuing with the experiment as described was sensible as there appears to be no simple experimental design that can accurately measure user ability to judge relevance against translation quality that is not confounded by other factors[1]. Therefore, we take the position of assuming that the additional factors in this experiment do not contribute significantly to our experimental results.

Therefore, following the iCLEF design, eight subjects were presented with retrieved documents from four queries in two different situations: half of the queries retrieved on the French collection and the other half on the English collection. A latin square design was used to ensure that query order and presentation of system did not confound the experiment. Users were given twenty minutes to judge the documents retrieved for each query. The subjects were Sheffield University students, who were native English speakers. They spent three hours in total on the experiment being paid £20 for their participation.

Results

Results from the initial data returned by iCLEF are shown in the table below. The effectiveness of users was determined using Van Rijsbergen's *F* measure (Van Rijsbergen 1979), where user judgements were compared to those made previously by assessors. The variable *a* was set to values of 0.2 and 0.8 to bias *F* to indicate user preference for recall and precision respectively.

System	***F*** (*a*=0.2)	***F*** (*a*=0.8)
Le Monde	0.49	0.60
LA Times	0.40	0.46

As can be seen for both values of *F* users judge relevance better on the translated French documents than on the English originals, however use of a t-test indicated that the differences were not significant. We believe that despite the potential problems with the experimental design, we have shown with some degree of confidence that the users reading the retrieved machine translated documents are more than able to judge the relevance of the retrieved text.

As described in the main iCLEF paper (Oard & Gonzalo Arroyo, 2001), users were asked to judge documents as relevant, not relevant, or somewhat relevant. The table above shows results of the user judgements focussing only on documents marked as relevant. The table below shows results re-calculated when documents marked as somewhat relevant are included. As can be seen, the difference in *F* values between Le Monde and LA Times is somewhat smaller particularly in the precision oriented *F* measure, indicating that users are more accurate in judging the relevance of marginally relevant native language documents than they are of translations.

System	***F*** (*a*=0.2)	***F*** (*a*=0.8)
Le Monde	0.65	0.59
LA Times	0.58	0.52

Finally, the degree of overlap between the sets of relevant documents (judged by the assessors) and the experimental subjects was measured. Overlap is defined as the intersection of the two sets divided by the union. In Voorhees's work (1998) overlap between pairs of assessors was found to range between 0.42 and 0.49. Taking the user judgements of those judged both relevant and somewhat relevant, the overlap ranged between 0.39 and 0.47, a similar range.

Conclusions and future work

In this report, we have described an experiment that compares user ability to judge relevance of documents written in different languages. The difficulty of designing such an experiment was discussed and the results of the experiment presented. The conclusion from the results was that for the documents tested here, French documents automatically translated into English using a good machine translation system are sufficiently readable to allow users to make accurate relevance judgements.

[1] One could design an experiment where native English speakers judge the translated French documents and native French speakers judge the same (un-translated) French documents. However, the two groups of users are likely to have different cultural backgrounds, which may influence the results. Whether this difference would influence experimental results more or less than the design chosen here can only be determined through further experimentation. Note, there was a very pragmatic reason for not pursuing this design: finding a sufficient number of French speakers would have been hard to achieve.

Extensions of this work would involve conducting further experiments to expand the number of users to try to find statistical significance in the data. In addition, exploring other experimental designs will also be a priority.

Acknowledgements

The first author conducted this work as part of her Master's course at Sheffield. Financial support for the work in this report was provided jointly by the MIND (IST-2000-26061, mind.cs.strath.ac.uk) and Clarity (IST-2000-25310, clarity.shef.ac.uk) projects.

References

Oard, D.W., Gonzalo Arroyo, J. (2001): The CLEF 2001 Interactive Track, in *Working notes of the Cross Language Evaluation Forum*

Van Rijsbergen, C.J. (1979): *Information retrieval* (second edition), Butterworths, London

Voorhees, E. (1998): Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness, in *Proceedings of the 21st annual international ACM-SIGIR conference on Research and development in information retrieval*: 315-323