Design and Evaluation of a Psychological Experiment on the Effectiveness of Document Summarisation for the Retrieval of Multilingual WWW Documents

Joanne Capstick, Gregor Erbach and Hans Uszkoreit
German Research Center for Artificial Intelligence GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
{capstick,erbach,uszkoreit}@dfki.de

Abstract

Since for the foreseeable future, retrieval will be an interactive task of the user looking through lists of potentially relevant documents, adequate support through various types of information is very important. A psychological experiment was conducted to examine the extent to which different types of automatically generated summaries aid retrieval and systematically evaluate user needs and behaviour in the area of cross-language retrieval for the WWW.

Motivation

The research described here was carried out in the MULINEX project (Erbach et al. 1997), the goal of which is to develop techniques for the effective retrieval of multilingual documents from the WWW. Special attention is given to presenting search results in such a way that users are supported in selecting documents that are relevant to their information needs. The presentation of search results includes: the language of the document; an automatically generated summary; a thematic classification; title; URL and document size. The purpose of the summaries is to increase retrieval effectiveness (recall of relevant documents) and efficiency (precision in visiting relevant documents without wasting time by looking at too many irrelevant ones). In order to get some information as to what extent different types of automatically generated summaries aid retrieval and systematically evaluate user needs and behaviour in this area, we conducted a psychological experiment. Subjects were given different kinds of summaries and their retrieval performance on two tasks were measured.

Design of the Experiment

Since the purpose of the summary is given by its application context we decided to use a mock-up system. This enabled us to simulate the search task in a realistic way while reducing variance as much as possible.

The subjects were given the task to submit a predetermined query to a search engine, and to select documents which were relevant to a given information need by looking through lists of potentially relevant documents. The information needs were formulated as follows:

1. What is good and bad for the heart?
2. What are the effects of ozone on human health?

A total of 84 subjects were tested; all of them were native speakers of German and they were principally humanity and law students. Six groups of 14 subjects were presented with different kinds of results lists. Documents were ordered by relevance, or theme:

A1. relevance, first n characters of document text
A2. relevance, query-independent summaries
A3. relevance, query-specific summaries
B1. theme, first n characters of document text
B2. theme, query-independent summaries
B3. theme, query-specific summaries

In group B the secondary ordering criterium was relevance. In group A the category of a document was stated, although this was not an ordering criterium. The result lists presented to the subjects contained 100 documents from three languages (English, German, French). The result lists were retrieved with AltaVista before the experiments in order to eliminate any variation that might arise by changes to the document base during the duration of the experiments and problems with dead links, moved pages etc. The thematic categories were assigned manually to the documents, and the summaries were generated automatically. The length of the summaries was fixed at 200 characters.

The query-independent summaries were generated by making use of the structural markup of documents (headings) and by layout specifications (boldface, font size, etc.). The query-specific summaries were generated by giving precedence to sentences and text segments in which stemmed query terms occurred.

Subjects were asked to use summaries and thematic categories to select documents relevant to the stated information need. In addition, they could optionally access machine translations of the summaries which were generated by Systran.

Data Gathered

Questionnaire Each questionnaire contained demographic information about the subject and feedback about aspects of the system being examined, i.e. summaries, summary translations, thematic categories.

Subjects were asked to make judgements about their language and computing skills for use as control data. Subjects also provided feedback about the appeal and usefulness of the user interface and the information presented. To obtain this feedback a semantic differential designed for software evaluation was used.[1] Subjects were also asked to provide additional comments.

Logfiles The logfiles contained the following information for each request to the web server: date and time; URI-encoded data and path information, e.g. documents and document summaries visited.

In addition, each document in the document base was assigned a human judgement of the relevance to the given information needs.

Evaluation of the Experiment

By analysing the logfiles we expected to be able to compare the effectiveness of different summary types and gain insight into the subjects' use of summary translations, thus giving us information about subject performance and behaviour. In addition, the evaluation of the semantic differentials from the questionnaire and the subjects' comments told us about the subjects' assessment of the system.

The quantitative evaluation of subject performance and the semantic differentials did not provide statistically significant results. However, the qualitative evaluation of the summaries and summary translations proved fruitful. The following is an overview of the results gained from the experiment. More details can be found in (Unz et al. 1998).

Subjects' Comments

Summary Type The comments are ordered according to positive and negative aspects, as well as areas for improvement. The most prominent comments for each summary type are listed.

[1] A semantic differential makes use of a graded rating scale which enables a quantitative evaluation of subjects' responses.

The availability of summaries was praised. They were seen as a time saving factor, enabling greater efficiency. Their compact representation and briefness was also positively commented upon.

A1.&B1. helpful, informative, time-saver.
A2.&B2. helpful, informative, time-saver; compact rep.
A3.&B3. informative, enable quick overview; briefness.

The lack of quality and clarity of summaries was criticised and subjects requested longer summaries (this was also suggested as an improvement by a few subjects). In the case of summaries consisting simply of the first 200 characters their abrupt termination within a word was criticised. A couple of users also commented on the fact that these were not real summaries rather the beginning of the document.

The query-dependent summaries were criticised by a few users as being disjointed – clearly the fault of our generation method – and being unrepresentative of the actual content of the page, which could often be the case with query-specific summaries, since the focus of the summary is aspects of the document relevant to the query and not the document as a whole.

A1.&B1. uninformative; too short; abrupt termination; not real summary.
A2.&B2. misleading; too short.
A3.&B3. uninformative; too short; lack of coherence; unrepresentative of document content.

Users from the different groups, although principally from the query-independent group, suggested listing the key words and terms from the text in the summary.

A1.&B1. extend summary.
A2.&B2. more precise with keywords.
A3.&B3. more informative.

Summary Translations These were used principally as an aid to understanding due to poor foreign language skills or simply out of interest and curiosity. A few subjects used the translations to test their own translation.

Reasons given for not using translations were lack of necessity and sufficient knowledge of the non-native languages, also that enough information was available in German. Some subjects discontinued the use of summary translations out of frustration due to the quality of the translations. A few subjects found the use of summary translations too time consuming.

Semantic Differentials

According to the semantic differential, there was no statistically significant effect of the summary type on the subjects' assessment of the summaries; although the query-independent summaries received a slightly higher rating than the others. The summary type used also had no statistically significant effect on the subjects' assessment of summary translations.

The semantic differential used was designed for software evaluation in general. There is no doubt a need to identify more appropriate factors for the user-rating of summaries in the retrieval context. These could be obtained in part from summary typologies as outlined by Hutchins (Hutchins 1993) or intrinsic criteria such as conciseness. The subjects' comments could prove a useful source of criteria, e.g. time-saving vs. time-consuming, coherent vs. disjointed.

Analysis of the Logfiles

62 subjects visited a German summary translation (12 of them only once). 21 subjects visited an English and 10 a French summary translation.

Effect of the Summary Type No statistically significant difference between subjects' performance[2] wrt. the summary type used was found.

This may be due to a number of reasons such as the quality of the summaries and the generation methods used. Perhaps the summaries were too short to be able to fulfil their purpose. The subjects were perhaps not placed under enough time pressure. The fact that subjects had to deal with summaries in different languages perhaps reduced the significance of the summary type. If the problem was lack of control, i.e. too many variable factors, then a *tighter* test environment needs to be designed. This could be achieved by simplifying the system and reducing variability, e.g. removing the multilingual dimension and only using German documents. However, it is important to ensure that the users' goals and the summary purpose are preserved. It is clear that we need a more appropriate or improved method for the quantitative evaluation of the summary types.

Conclusion

In spite of criticisms regarding the quality and clarity of the summaries, subjects still regarded them as a helpful and time-saving factor for the task in hand. Subjects would, in general, prefer longer summaries; although this may lead to screen real-estate problems. Subjects also praised the availability of summary translations, although quality was a problem here too. From the comments, it seems necessary to reduce the time and effort needed to use summary translations.

[2] Performance was measured as the number of relevant documents visited.

We have not yet found the right way to test the summaries quantitatively; although the experiments give indications as to which methods may work better in the future. There is the need to set up a more controlled test environment which does not disconnect the summary from its purpose or change the user goals. The comments were rendered especially valuable by the fact that they were made about a concrete situation just experienced by the subjects, and not disconnected from the situation as is often the case with questionnaires. The comments could also serve as a useful source of ranking criteria for semantic differentials used in future summary evaluation work.

A possible extension to this work may be the in depth analysis of summaries for cross-lingual information retrieval wrt. a descriptive framework for summaries (e.g. Spärck Jones 1993); and the identification of mappings between elements of the framework and evaluative methods.

Acknowledgments

MULINEX is funded by the European Commission's Telematics Applications Programme (Language Engineering Sector, LE-4203).

The experiments were designed, carried out and evaluated by DFKI and the media psychology institute, MEFIS e.V.

References

Endres-Niggemeyer, B., Hobbs, J., and Spärck Jones, K. eds. 1993 Workshop on Summarizing Text for Intelligent Communication. Dagstuhl.

Erbach, G., Neumann, G., and Uszkoreit, H. 1997. MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World Wide Web. In Hull, D. and Oard, D. eds. *Cross-Language Text and Speech Retrieval – Papers from the 1997 AAAI Spring Symposium.* AAAI Press, Stanford.

Hutchins, J. 1993. Introduction to Text Summarization Workshop. In Endres-Niggemeyer, B., Hobbs, J., and Spärck Jones, K. eds. *Workshop on Summarizing Text for Intelligent Communication.* Dagstuhl.

Spärck Jones, K. 1993. Summarising: Analytic Framework, Key Component, Experimental Method. In Endres-Niggemeyer, B., Hobbs, J., and Spärck Jones, K. eds. *Workshop on Summarizing Text for Intelligent Communication.* Dagstuhl.

Unz, D., Capstick, J., Erbach, G. Heidinger, V., and Uszkoreit, H. 1998. Psychological Experiment on the Presentation of Search Results for the Effective Retrieval of Multilingual Documents from the WWW. Technical Report, DFKI, Saarbrücken.