ExtraWeb
An Extrinsic Task-oriented Evaluation of Webpage Extracts

Patrick Pedreira Silva[1] and Lucia Helena Machado Rino[2]
[1]*Centro de Ciências Exatas e Sociais Aplicadas, Universidade Sagrado Coração, Rua Irmã Arminda, Bauru, SP, Brazil*
[2]*Departamento de Computação, Universidade Federal de São Carlos, São Carlos, SP, Brazil*

Keywords: Webpage Extracts, Automatic Summarization, NLP, Ontology.

Abstract: This paper focuses on the usefulness of extracts of webpages in Brazilian Portuguese as the means to filter information for Web users to quickly and consistently judge the relevance of search engine results. ExtraWeb, an ontology- and HTML-based summarizer, has been built aiming at providing an alternative to query-biased extracts typically made available by Web search engines. An extrinsic evaluation of ExtraWeb was carried out under a controlled experiment that retrieves webpages in Portuguese and generates a set of extracts for an Internet user to evaluate. Only the relevance judgment of the extracts was assessed. Results show that the system is promising in helping users to filter relevant webpages.

1 INTRODUCTION

In the last decade, the exponential increase of the amount of available online documents has brought up significant importance to techniques of Automatic Summarization (AS). This is mainly due to the information overload: the huge amount of documents poses to the user a difficult task to locate relevant information. Users can find it hard to judge the relevance of retrieved documents and, thus, to satisfy their information needs based only on a snippet that contains common words to their queries. Such a wording, very often, does not help the user finding relevant documents. Aiming at overcoming this, we built a system, henceforth ExtraWeb, which applies ontology- and HTML-based summarization techniques to identify and extract the main excerpts of documents available in the Internet. By using HTML tags and concepts provided by the ontology, ExtraWeb focuses on semantic processing for producing a more useful extract than just the usual Web snippets.

Ultimately, ExtraWeb aims at improving the user satisfaction by providing her/him with the means to filter relevant documents from a collection by simply reading the extracts. The cognitive overhead of the user, usually implied by most search engine interactions, can be thus minimized (Conklin, 1987).

Issues concerning indexing and searching as such are disregarded here. Hence, AS for Brazilian Portuguese (BP) was explored apart from a Web engine. However, our modelling corpora – of Web documents – and ExtraWeb assessment were provided under a real setting: we adopted Google strategies when needed, as we shall see in Section 3.

In what follows, we first introduce ExtraWeb, along with the motivation of the present work (Section 2). Then, we describe the extrinsic evaluation carried out (Section 3) in an information retrieval scenario, which yields our discussion in Section 4. Relevant issues are discussed in Section 5.

2 ExtraWeb: A WEBPAGE SUMMARIZER

Most search engines usually present the user a description that literally reproduces retrieved document content. Very often, such a description conveys few and obscure, or non-relevant information for the user to decide whether to go after the corresponding document. Moreover, descriptions usually refer to the titles of the available documents (White et al., 2002).

Aiming at making the descriptions of a search engine clearer and more significant to the user, ExtraWeb is used to produce more useful extracts of the retrieved documents. It takes into account

Pedreira Silva P. and Machado Rino L..
ExtraWeb - An Extrinsic Task-oriented Evaluation of Webpage Extracts.
DOI: 10.5220/0004449104670474
In *Proceedings of the 15th International Conference on Enterprise Information Systems* (ICEIS-2013), pages 467-474
ISBN: 978-989-8565-59-4

HTML tags and ontological information to rank document units for their relevance to compose the extracts. Two types of information are used here: keywords, or words of a document that are attached to special HTML tags, and the main topics of the document, which are delineated through the ontology. HTML tags considered for recognizing keywords must be previously introduced by the very Webpage author. Supposedly s/he uses them for emphasis and, thus, for pinpointing relevant information. Stylistic tags such as the emphasis one (<I></I>) and bold (<B></B>) are examples.

Besides being relevant, keywords actually also signal topics considered relevant, now by the very author of the document. So, ExtraWeb aggregates two distinct accounts on determining relevant topics to include in extracts: one provided by the author of Web documents, and another, by the system reasoning on ontological concepts.

To determine the main topics of a document, ExtraWeb uses the Yahoo Ontology for BP, which has been manually enriched, to yield a more fine-grained conceptual representation of the information conveyed by webpage documents. Following work by (Lin, 1995) and (Tiun et al., 2001), we collected lexical items from a modelling corpus and used a thesaurus (Greghi et al., 2002) for synonymic relations amongst the items. Only nouns, verbs, and adjectives were considered. For each lexical item, we retrieved its position in the ontology. Whenever it did not match directly a concept, we identified its immediate vicinity (a parent-child context) concept and included it accordingly in the hierarchy. Extra leaves were the simplest case of enrichment, such as in including the child concept corresponding to the lexical item 'Dodge' in the "Automotive/ Classic_Cars" relationship, yielding "Automotive/ Classic_Cars/ Dodge". Yahoo Ontology for the BP language amounts to c.a. 5.5 thousand concepts. Our enrichment targeted about half of it, resulting in c.a. 26 thousand concepts. In producing this refined version of the ontology, mapping significant words onto concepts was simplified during topic identification.

In determining relevant information to summarize a document, using keywords is a classical approach (Luhn, 1958); (Edmundson, 1969). Considering ontological concepts for topic identification has been suggested by others. Usually, the probability of a document fragment (a lexical item or phrase) to be related to a node in the ontology is calculated (Mladenic, 1998). The semantic relationship between documents keywords and concepts in the ontology is also used in AS (Tiun et al., 2001). In any case, keyword- or ontology-based methods can be used to determine highly classified concepts matching content words that signal the main topics of a document, to include in its extract. Keywords can also be used in isolation to classify the sentences that embed them.

In ExtraWeb, we considered the use of keywords separately from the use of the ontology, as complementary strategies. Once identified the main topics of a document through either method, the corresponding document units were ranked and normalized, yielding their relevance degrees. The highly classified ones were, thus, chosen to compose an extract in the usual way, i.e., according to a given compression rate.

3 EXTRINSIC TASK-ORIENTED ASSESSMENT

According to (Dorr et al, 2005) the usefulness of a summary can be measured through an extrinsic task of relevance judgment. Judges are asked to compare their judgments on extracts with their own judgments of the corresponding full documents. Dorr et al (2005) claim that such an assessment in the IR context is more reliable than other gold-standard measures, such as the LDC-Agreement method used in SUMMAC (Mani et al., 2002). We similarly adopt this methodology in ExtraWeb assessment. Extracts of documents resulting from a direct access to websites were presented to a judge committee, composed of usual Web users. In our context, usefulness is meant to apply to those extracts that are clear enough for the Web user to decide if the hidden document is of interest for him/her, to follow the link provided by a search engine. In being useful, we also consider them to be relevant to the task.

Tests for statistical significance of the judges' answers were also provided, based upon the p value (Mann-Whitney test). Only when p < 5% data comparison is statistically significant. We also adopted a similar setting for evaluation to (Amitay, 2001) and (White et al., 2002), which emulates a user session in an Information Retrieval context, as described below.

3.1 The Design of the Experiment

The following issues were considered important: typical queries usually are short, c.a. 1 to 3 words long (Jansen et al., 2000); (Inktomi-Corp, 2003);

few answers of a search engine are actually relevant, from all the retrieved ones (White et al., 2002); Web users most often need to single out the relevant answers, in order to retrieve the documents that are most suggestive to them.

Following the above, the experiment was conducted in such a way as to allow a user to access answers through extracts. For simplicity, in this paper we do not distinguish simple fragments from real sentence-based extracts, although the former ones may be totally non-textual (i.e., only a snippet) and the latter may be quite cohesive and coherent.

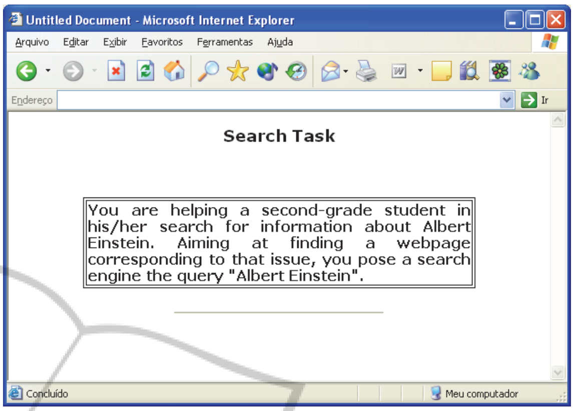
For each document and query, three distinct extracts were produced, which were generated by three hidden summarizers: ExtraWeb, Google, and the Baseline. Google was chosen because it is widely used and considered by many as the best search engine (Griesbaum, 2004). It very often produces only query-biased snippets that have no compromise with coherence. The Baseline summarizer just collects document headlines. Both systems usually show 1-to-2-lines long extracts. Source documents were obtained through Google previously to the evaluation task. Only the top three were considered in the assessment, under the assumption that those are the most relevant ones. After showing the extracts to the judges, their task was to compare them and fill in a questionnaire, as explained next.

3.2 The Oriented Task Itself

Fifty eight subjects were invited to contribute with their judgments by going into a webpage of the experiment. Firstly, a description of a task was shown, to make clear what it was emulating (Fig. 1). Five tasks were defined based upon five distinct queries. The one supplied to a judge was randomly selected, delineating an individual session. This was considered to fulfil the experiment only if completely executed.

The task itself was partially fixed in advance: queries were not allowed to be user-defined, to ensure that judgment was uniformly applied to consistent sets of results, one for each summarizer under consideration. Consistent result sets are only those coupled with their corresponding queries (Amitay, 2001).

The readership, or judging, community, was set free: the experiment was made widely available through a browser, and judges' tasks and answers were recorded into log files. Other relevant information could also be assessed through those files, such as the judges' identification (58 IP numbers), to verify if they had duplicated their judgment. Also free was the retrieval of documents by Google under any given query.



Search Task

You are helping a second-grade student in his/her search for information about Albert Einstein. Aiming at finding a webpage corresponding to that issue, you pose a search engine the query "Albert Einstein".

Figure 1: Screen display of a query-bound task.

Subjective judgment was carried out as follows: once the judge was presented a task, a collection of three document links and extracts was shown, for each summarizer. However, ExtraWeb and Baseline extracts were conditioned to the previous retrieval of documents by Google. In other words, once Google generated its own extracts and document links, these were used by the other systems to produce their corresponding extracts. The judge was then asked to read the extracts and select only the one s/he found most appropriate for the query. Then, s/he was asked to click on the link by that extract, in order to see the document webpage. After reading the document, the judge had to fill in the online questionnaire form.

The above steps were repeated for the results of the three systems and the recorded judgment sessions were later compiled (Section 4).

3.3 Extracts Generation

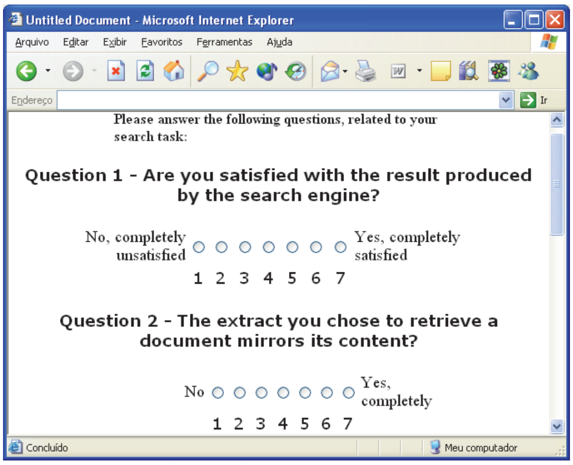
Omitting titles and URLs of the documents was a crucial constraint to automatically generate the corpora of extracts, to prevent spoiling the experiment. This can happen when titles or URLs somehow substitute the extracts: many people do not realize that extracts provided by search engines are faulty or non-informative due to the existence of titles or URLs that come along the document in the collection (Mani et al., 2002).

3.4 The Questionnaire

Following (Amitay, 2001); (Lewis, 1995); (Perlman, 2011), to build the questionnaire we took care of its preciseness and brevity: users want to quickly fill it in and proceed to the conclusion of the experiment. Five questions were, thus, chosen, to avoid an excessive load on the judges. They are mostly translations into Brazilian Portuguese, of Amitay's questionnaire in English for the similar task. We adopted such strategy because they are very general and mirror usual queries proposed on the Web. Proper names of famous people, leisure activities, research interests, health issues and other factual information are examples of real queries.

Screen dumps of the form to fill in are shown in figures 2 and 3 (in English only for practical purposes). It is worthy noticing that such queries could be anything but too sophisticated terms, which would address a more knowledgeable readership. In turn, evaluation and answering the questionnaire would result too difficult for the judges.

The questions focus upon diverse assumptions. The first two (Fig. 2), e.g., aimed at assessing the satisfaction of the judge with the usefulness of his/her chosen extract and its correspondence with the respective Web document. This coincides with the main goal of any search engine. Questions 3 and 4 in Fig. 3, instead, focus on the judges' profiles, mainly, on their regularity during the assessment. Question 3 refers to just the assessment accomplishment, whilst Question 4 is very broad. Answers to them helped us detect whether the judge chose her/his extracts actually based on comparing them. Such questions are, thus, content-independent. Assuming that the answer to Question 3 was option 1, the last question aimed at an overall evaluation of the automatic extracts shown on the screen for each summarizer and task. Apart from questions 3 and 4, the others actually aimed at measuring only the quality of the extracts with respect to the decision making of Web users.


Untitled Document - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Endereço
Ir
Concluído
Meu computador

Please answer the following questions, related to your search task:

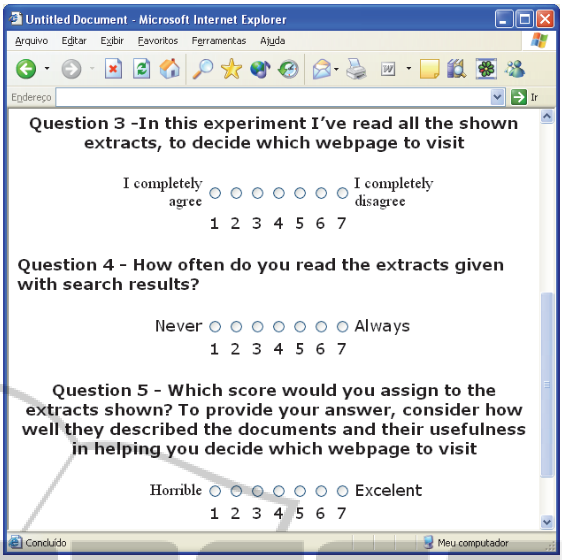
Question 1 - Are you satisfied with the result produced by the search engine?

No, completely unsatisfied ○ ○ ○ ○ ○ ○ ○ Yes, completely satisfied
1 2 3 4 5 6 7

Question 2 - The extract you chose to retrieve a document mirrors its content?

No ○ ○ ○ ○ ○ ○ ○ Yes, completely
1 2 3 4 5 6 7

Figure 2: Questions 1 & 2.


Untitled Document - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Endereço
Ir
Concluído
Meu computador

Question 3 -In this experiment I've read all the shown extracts, to decide which webpage to visit

I completely agree ○ ○ ○ ○ ○ ○ ○ I completely disagree
1 2 3 4 5 6 7

Question 4 - How often do you read the extracts given with search results?

Never ○ ○ ○ ○ ○ ○ ○ Always
1 2 3 4 5 6 7

Question 5 - Which score would you assign to the extracts shown? To provide your answer, consider how well they described the documents and their usefulness in helping you decide which webpage to visit

Horrible ○ ○ ○ ○ ○ ○ ○ Excelent
1 2 3 4 5 6 7

Figure 3: Questions 3 to 5.

A judge could grade a question by ticking only one out of the seven provided options. The 7-point scale was suggested by (Tang et al., 1999) for its utility in measuring the confidence of the users' judgments. The following queries are fixed in ExtraWeb assessment (hereafter identified by Qi, i varying from 1 to 5): Albert Einstein, *Projeto Genoma* (Genome Project), origami, *vida selvagem* (wildlife), *fusos horários* (world time zones).

4 RESULTS OF THE VOLUNTEER JUDGMENT

To compile the judges' answers, our first task was to retrieve their logs to verify if each of them consistently answered the questionnaire. Also, we verified if they correctly followed the instructions, i.e., if they thoroughly answered all the questions in a single interaction and interacted with each system only once. If there were more than one user interaction originating from the same IP number, only the first one was taken. Interactions considered undesirable were those that did not convey all the answers to the five questions or those that went back and forth, breaking the normal proposed sequence for answering. Violation of such conditions implied suppressing that log file from the judges' answer corpus. After such a filtering, only 169 out of 290 answers remained valid for analysing ExtraWeb performance. Table 1 shows the answers distribution across the different tasks and sets of extracts.

Table 1: Distribution of the judges' answers.

	Total	Search Tasks				
		Q1	Q2	Q3	Q4	Q5
Google	60	9	19	11	13	8
ExtraWeb	56	9	16	10	13	8
Baseline	53	9	16	10	13	5
Total	169	27	51	31	39	21

As we can see, answers were quite uniformly distributed amongst the three systems. This is important to assure a balanced comparison. Particularly, users interacted many times with tasks 2 and 4.

Observing the results for Question 5 (Fig. 4), there was not a significant difference between Google (a 5.4 av. score) and ExtraWeb (a 5.2 av. score) performances. These averages are not statistically significant either – their p value equals c.a. 35%, well above the 5% usual limit. Having ExtraWeb extracts as good as Google ones could make us question the use of our new proposal. However, for one of the tasks (Task 4), ExtraWeb outperformed Google, as shows Fig. 5. This signals that the system usefulness may be query-dependent and, thus, it deserves further investigation.

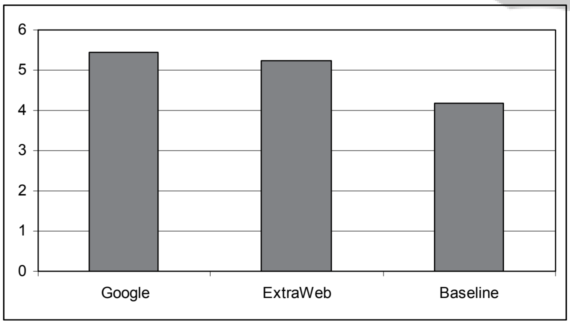


Figure 4: Mean values for Question.

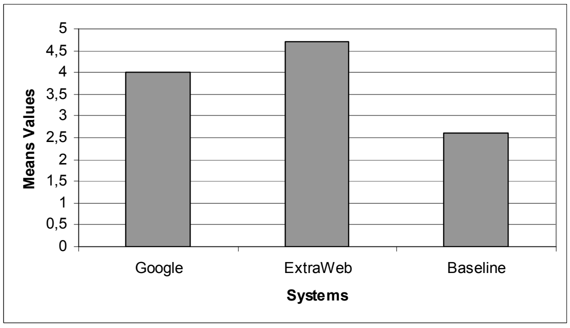


Figure 5: Mean values for Question 5 (Task 4).

From the total of 58 judges who compared at least those two systems, 58% (34) scored the same or higher for ExtraWeb extracts in Question 5. So, our extracts may be significantly useful for the user to judge the relevance of the corresponding documents. Scores were also most often produced on all the extracts shown to the judges (see average scores of Question 3 in Table 2), despite the fact that only in c.a. 83% of the interactions judges said they read extracts produced by search engines (Question 4).

Table 2: Average answer scores.

Question	Google	ExtraWeb	Baseline
1	5.3	4.8	5.0
2	5.6	5.2	4.7
3	6.7	6.6	6.4
4	5.8	5.8	5.7
5	5.4	5.2	4.2

Table2 shows absolute values in the 7-point scale. Percentages were calculated by normalizing those in the 100 scale. Table 2 also shows that the judgments were quite consistent: Questions 1 and 5 were averaged about the same for Google and ExtraWeb. Considering Question 1 in isolation, ExtraWeb was scored low (68% for full satisfaction only). This may be explained by the fact that users were not able to pose their own queries and even the task was provided in an artificial setting. In a real scenario, though, satisfaction may increase. Oppositely, non-satisfaction with the results for a given task could be due to a poor performance of ExtraWeb due to other factors, amongst which: (a) writers do not use HTML consistently or broadly, as already referred to. In other words, the documents may be poorly annotated in HTML, yielding such a weak Web document structure that would not allow identifying relevant fragments. (b) the language of the Web documents could be so poor (e.g., conveying spelling errors or slangs) that the correspondence with our ontological model would be hampered.

Scores for Question 2 show that in 80% of the cases Google extracts were considered to mirror the documents content, against 74% of the ExtraWeb ones. This difference is not statistically significant (p=22%), but such performance could be improved if HTML tags were more appropriately used. Answering Questions 3 and 4, users also claimed to have read the three types of extracts with no significant differences (p=36% and 83%, respectively, for Google and ExtraWeb).

Actually, the slight better performance of Google, when compared to ExtraWeb, may simply be due to the fact that ExtraWeb works on documents previously collected by Google itself. In this case, Google acts just as a search engine and not as a summarizer. As a consequence, ExtraWeb may accumulate the fragilities of both systems. Another reason may be an inadequacy on mapping lexical items on the concepts of the extended Yahoo ontology. None of the above issues have been deeply explored.

Also worthwhile considering is the way information conveyed by the extracts influenced the users' analysis of the results. The presentation of results out of a real search context could force users to read the extracts more carefully than usual. At the same time that this aimed at keeping the judges focused on the extracts, thus, on the goal of the experiment, it could prevent them to arrive at ideal choices for their answers.

5 RELEVANT ISSUES

In the last decade, the exponential increase of the amount of available online documents has brought up significant importance to techniques of Automatic Summarization (AS). This is mainly due to the information overload: the huge amount of documents poses to the user a difficult task to locate relevant information. Users can find it hard to judge the relevance of retrieved documents and, thus, to satisfy their information needs based only on a snippet that contains common words to their queries. Such a wording, very often, does not help the user finding relevant documents. Aiming at overcoming this, we built a system, henceforth ExtraWeb, which applies ontology- and HTML-based summarization techniques to identify and extract the main excerpts of documents available in the Internet. By using HTML tags and concepts provided by the ontology, ExtraWeb focuses on semantic processing for producing a more useful extract than just the usual Web snippets. Keyword- and topic-based methods for summarizing Web documents have been significantly explored lately for improving, speeding up, and making Web content, in general, more accessible to most people. Liang et al (2004) compare multi-word keyterm-based summaries with keyword-based ones, and show that the former are more effective in making Web content accessible to users. They select features to produce keywords as headline-like summaries. Relevant features to them are people's names, groups, events, and places. The results of their specific DUC 2004 task show that using feature selection does better than combining ranked words or other simpler methods such as location, to yield non-textual summaries. Even PageRank (Page et al, 1998) has been adapted for a content-based approach (Haveliwala, 2002) that computes the similarity of a query to topics. In this case, however, topics are retrieved from the Open Directory Project, instead of being directly related to the intended document collection.

In a diverse approach, Chirita et al., (2005) advocate in favour of personalizing search through metadata that mirrors documents topics and importance. Such type of data also helped building the Yahoo ontology, which aims at accessing content on the Web with high quality. Also using metadata and ontologies, Barros et al., (1998) focus on enhancing Web searches by considering information that can better contextualize the queries. Information, in this case, is dynamically provided by the ontologies, which act as providers of potential related words. The user may interfere in the process of contextualizing queries by selecting concepts that better help expanding the query. By searching documents through such expanded queries, precision and recall are claimed to be improved.

Our content-based approach in ExtraWeb relates to the above, in the search for more effective results of search engines based on content information. Mainly, it brings about the need to consider the relationship between query and topics conveyed either explicitly, by a keyword list given by the very author of a document, or implicitly, through the mapping of content words onto the Yahoo concepts. Using the Yahoo ontology in BP is quite adequate because it is a broad repository of electronically available information. Due to the lack of a BP WordNet we decided to enrich it, in order to make concept classification easier. However, it has not been fully populated with new concepts extracted from our modelling corpus. Given that ExtraWeb performance was considered as good as Google one, scaling up the amount of concepts in the ontology is very likely to yield better extracts for the users.

In some aspects, our evaluation is also similar to some of the DUC tasks, e.g., by being user-oriented and aiming at verifying if summaries fulfil the need for information expressed by a topic query. An example task on multi-document AS on DUC is that comprising a user profile, a DUC topic, and a cluster of documents relevant to the topic. Automatic summarizers should create a brief, well-organized, and fluent summary of the full cluster, satisfying the topic query. Summary evaluation was manual, on a 5-point scale (A= the best; E=the worst) for the well-formedness (i.e., readability and fluency) of the summaries. To assess judgment, five quality questions (against three, in our case) were used. The relative responsiveness (5-point scale, 1= worst, 5= best) of the summaries was also assessed by measuring the amount of information that helps the user in successfully retrieving information. This measure seems to coincide with our usefulness one, which is assessed in our experiment through questions 1, 2, and 5. The average responsiveness in DUC 2005 was 48% for automatic summaries, against 93% for the reference ones (Hachey et al., 2005).

Accordingly looking at our data, usefulness averaged 72% for those three questions. Certainly, this is quite a simplistic comparison, due to the profound differences of both assessments. However, the whole design of the experiment is quite significant, when compared to the DUC ones. To make it statistically significant, we must invest on its robustness (e.g., by increasing the amount of Web users and search engine answers).

6 FINAL REMARKS

The reported results show a significant proximity of ExtraWeb with Google. This means that ExtraWeb may also be useful for the users to make decisions on retrieving documents, in spite of their low score (68%) on full satisfaction with the results of the emulated search task. Although the experiment was not intended to control either the homogeneity of the judging population or its subjectivity in accomplishing the demanded task, the analysis of their scores shows that the overall judgment was quite consistent. However, the same extrinsic task-oriented evaluation may yield different results when a higher scale on both, judges and retrieved documents, is taken into account. Usually, users assessing the same task and set of results of a search engine would not necessarily respond in the same way and some of them might read the extracts more carefully than others. As a consequence, their judgments could be more accurate. This is very likely to be evident when scaling up the type of assessment reported in this paper.

Another important issue to pinpoint is that ExtraWeb is domain-independent. However, it depends on previous HTML-marking keywords which are usually accomplished by the documents authors. The alternative to this would be to generate a keywords list through statistical methods such as Luhn’s itself. However, this would not yield keywords as expressive as the authored ones.

Future work shall build on both, improving the enrichment of the ontology and assessing more broadly the system, in a distributed environment in real-time. Most probably, it will be relevant to reproduce similar quality questions to the ones used in the most recent DUCs too.

REFERENCES

Amitay, E., 2001. *What lays in the layout: Using anchor-paragraph arrangements to extract descriptions of Web documents*. PhD Thesis. Department Mani of Computing, Macquarie University.

Barros, F. A., Gonçalves, P. F., Santos, T. L. V. L., 1998. Providing Context to Web Searches: The Use of Ontologies to Enhance Search Engine's Accuracy. *Journal of the Brazilian Computer Society*, 5(2):45-55.

Chirita, P. A., Nejdl, W., Paiu, R., Kohlschütter, C., 2005. Using ODP meta-data to personalize search. In the *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178-185.

Conklin, J., 1987. Hypertext: An Introduction and Survey. *IEEE Computer*, 20(9), pp.17-41.

Dorr, B., Monz, C., President, S., Schwartz, R., Zajic, D., 2005. A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate? In the *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 1-8.

Edmundson, H. P., 1969. *New Methods in Automatic Extracting*. Journal of the ACM, 16(2):264-285.

Greghi, J. G., Martins, R. T., Nunes, M. G. V., 2002. Diadorim: a lexical database for brazilian portuguese. In the *Proc. of the Third International Conference on language Resources and Evaluation*. 4:1346-1350.

Griesbaum, J., 2004. Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. *Information Research*, 9(4), paper 189.

Hachey, B., Murray, G., Reitter, D., 2005. Embra System at DUC 2005: *Query-oriented multi-document summarization with a very large latent semantic*. Document Understanding Conference 2005, Vancouver, British Columbia, Canada.

Haveliwala, T. H., 2002. Topic-sensitive PageRank. *In the Proc. of the Eleventh International World Wide Web Conference*, Honolulu, Hawaii.

Inktomi-Corp., 2003. *Web search relevance test*. Ve-ritest. Available at http://www.veritest.com/clients/reports/inktomi/inktomi_Web_search_test.pdf [March 2006].

Jansen, B. J., Spink, A., Saracevic, T., 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207-227.

Lewis, J. R., 1995. Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction*, 7(1):57-78.

Liang, S. F., Devlin, S., Tait, J., 2004. *Feature Selection for Summarising: The Sunderland DUC 2004*

Experience. Document Understanding Conference 2004, Boston, USA.

Lin, C. Y., 1995. Knowledge-Based Automatic Topic Identification*. In the Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 308-310.

Luhn, H. P., 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159-165.

Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., Sundheim, B., 2002. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43-68.

Mladenic, D., 1998. Turning Yahoo into an Automatic Web-Page Classifier. In the *Proc. of the 13th European Conference on Artificial Intelligence (ECAI'98)*, pp. 473-474.

Page, L., Brin, S., Motwani, R., Winograd, T., 1998. *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. Rep., Stanford University, Stanford, CA.

Perlman, G., 2011. *Web-Based User Interface Evaluation with Questionnaires*. Available at http://www.acm.org/~perlman/question.html [January 2013].

Tang, R., Shaw, W. M., Vevea, J. L., 1999. Towards the identification of the optimal numbers of relevance categories. *Journal of the American Society for Information Science (JASIS)*, 50(3):254-264.

Tiun, S., Abdullah, R., Kong, T. E., 2001. Automatic Topic Identification Using Ontology Hierarchy. In the *Proc. of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2001)*, pp. 444-453.

White, R., Ruthven, I., Jose, J. M., 2002. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In the *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pp. 57-64.