Browsing for Information by Highlighting Automatically Generated Annotations: A User Study and Evaluation

Victoria Uren, Enrico Motta,

Martin Dzbor

Knowledge Media Institute, Open University
Milton Keynes, UK
(v.s.uren,e.motta,m.dzbor)@open.ac.uk

Philipp Cimiano

AIFB, University of Karlsruhe
Karlsruhe, Germany
cimiano@aifb.uni-karlsruhe.de

ABSTRACT

The realization of the Semantic Web is constrained by a knowledge acquisition bottleneck, i.e. the problem of how to add RDF mark-up to the millions of ordinary web pages that already exist. Information Extraction (IE) has been proposed as a solution to the annotation bottleneck. In the task based evaluation reported here, we compared the performance of users without access to annotation, users working with annotations which had been produced from manually constructed knowledge bases, and users working with annotations augmented using IE. We looked at retrieval performance, overlap between retrieved items and the two sets of annotations, and usage of annotation options. Automatically generated annotations were found to add value to the browsing experience in the scenario investigated.

Categories and Subject Descriptors

H.3.3 Information Storage and Retrieval: *Information Search and Retrieval – search process.*

General Terms

Performance, Experimentation.

Keywords

Knowledge Management, Semantic Web, Annotation, User Studies.

INTRODUCTION

The vision of the Semantic Web presupposes the existence of Web pages with semantic markup (annotations) based on shared ontologies. As a consequence, the realization of the Semantic Web lies on the far side of a knowledge acquisition bottleneck. Natural Language Processing (NLP) methods, such as Information Extraction (IE) and Named Entity Recognition (NER) have been proposed as solutions that could generate annotations automatically on a large scale, e.g. [6][11][3]. We support this proposal and are active in developing automated and semi-automated systems for the creation of semantic annotations [12][1].

If the Semantic Web is to rely on automatic mark up then robust evaluation will be required. IE systems have traditionally been evaluated using the kind of comparative, quantitative methods developed for the MUC conferences[1]. These methods have encouraged the development of effective algorithms. However, in this paper we present a complementary approach to evaluation, which looks at the user experience of working with systems partially populated using IE technology rather than the empirical performance of the underlying algorithms on a given dataset. Refinement of the methods may be required but we believe that this initial experiment explores another important facet of performance, i.e. whether the annotations produced by automatic annotation are fit for the task they were intended for.

The paper is structured as follows. We first briefly describe the technologies that were integrated in the evaluated system. Then we outline the methodology of the user study. We examine the effects of using a lexicon boosted using IE based annotation, on retrieval performance, answer coverage between retrieved results and annotations, and the users interaction with annotation options and their perceptions of the system as a whole. We conclude that IE based annotation can enhance Semantic Web systems by increasing both the quantity and scope of annotations.

TECHNOLOGIES

This evaluation looked at whether two NLP tools, PANKOW [1] and ESpotter [14] could be used to successfully produce annotations which could be highlighted with the Magpie semantic browser [7] in order to enhance search performance. They were used to construct one of the lexicons from which Magpie generates semantic annotation on the fly. In the context of our evaluation changing these lexicons provides a means to apply different annotation schemes to the same underlying data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
K-CAP'05, October 2–5, 2005, Banff, Alberta, Canada.
Copyright 2005 ACM 1-59593-163-5/05/0010...$5.00.

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_data/muc_data_index.html

Magpie

Magpie [7] is a framework developed by The Open University partially responding to the challenge of the knowledge acquisition bottleneck. It allows users of web-based documents to interpret content from different conceptual perspectives by automatically generating annotations corresponding to a particular ontology as a semantic layer over the content of the document. This allows Magpie to provide semantic web services for documents with no semantic mark up, or which are marked up according to ontologies that do not suit the user's purpose.

The end-user part of the Magpie framework comprises a browser plug-in (currently available for Microsoft Internet Explorer or Mozilla). The plug-in enables the user to choose an ontology and to toggle categories of knowledge via simple push buttons presented in a toolbar (see Figure 1(A)). Selecting a button highlights items in the text that are relevant to the chosen category. The user can access a menu with relevant functionalities for each annotated item.

These dynamic annotations are generated using a lexicon which relates each concept in the ontology to the various text strings by which it is commonly represented. Previously, Magpie lexicons were constructed by domain experts. We have tried to automate this process in this experiment using the two IE tools PANKOW and ESpotter.

PANKOW

PANKOW is a web-based information extraction system developed at the University of Karlsruhe. The system is based on the assumption that semantics in general and annotation in particular can be approximated by examining the distribution of certain syntactic patterns conveying the relation of interest for the item in question (compare [1]). PANKOW first identifies proper nouns, e.g. "Magpie", using a part of speech tagger, and then searches the Web for syntactic patterns which provide evidence that the proper noun belongs to one of the classes of the ontology. For example it searches for definite expressions like "the Magpie project" to associate "Magpie" with the class "project". In the context of the experiments reported in this paper we were concerned with formal annotations of instances appearing in the KMi stories dataset.

We processed 307 KMi planet stories with the PANKOW system as described in [2]. Overall PANKOW yielded 1270 annotations (4.1 per document) which on a scale from 0 to 3 were rated with 1.8 credits on average by a human evaluator. If we regard every annotation receiving at least 2 credits as correct this translates into an accuracy of 58%. A total of 755 entities whose PANKOW classifications were recognized as belonging to one of the nine AKT++ upper level categories were added to the lexicon.

ESpotter

ESpotter, is a named entity recognition (NER) system developed by the Open University [14]. It builds on the basis of standard named entity recognition (NER) methods, such as patterns (which exploit features such as capitalization of people's names) and lexicons (for example lists of common names) but also incorporates a domain adaptation mechanism which allows it to choose the methods which are most likely to be reliable for a particular site. Given a Web page and its URI, ESpotter pre-processes it by removing mark-up tags etc., finds regular expressions, which have high probabilities for NER on the domain, and uses them to recognize entities of various types on the page.

ESpotter extracted a total of 761 annotations (approx. 2.4 per document) from the KMi Planet News stories. These were 428 entities found for Organization, 243 for Person, 4 for Research Area and 86 for Project.

A separate study of the extraction performance of ESpotter [14] suggests that on the KMi Planet News stories it has recall results above 85% for the categories of extracted data used in this experiment and precision values above 90%. This high accuracy is a result of ESpotter's adaptation mechanism which has been fine tuned for the KMi Portal.

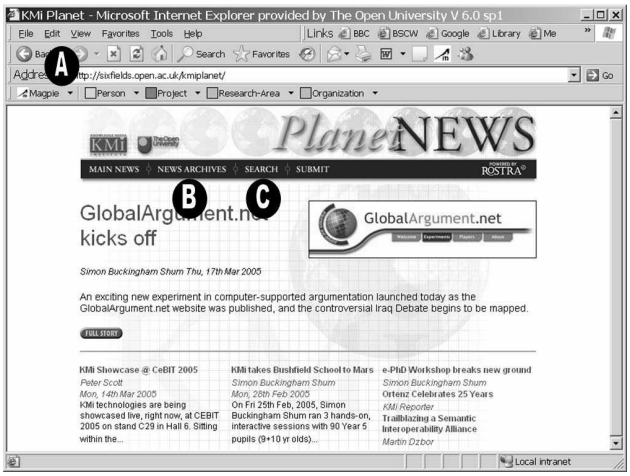

Figure 1 The Planet News interface as seen by Group B (Magpie/AKT) featuring (A) Magpie highlighting options, (B) News Archive, (C) Search option.

EVALUATION METHOD

The evaluation took the form of a user study and was conducted jointly by researchers from the Open University and University of Karlsruhe at KMi in November 2004. Our aim was to see whether or not semantic annotations generated by IE improved the performance and experience of Magpie users on information gathering tasks. The performances of three groups of participants were compared on two fact retrieval tasks which involved searching an online database of news stories. The Groups were: Group A (baseline), who used only the news stories, Group B (Magpie/AKT) who had the news stories and a version of Magpie with a hand crafted lexicon based on internal knowledge bases from KMi and the University of Southampton, and Group C (Magpie/AKT++) who had the same set up as Group B but with the hand crafted lexicon enhanced by

additions from the information extraction tools and additions of recent information from KMi's knowledge bases that had been created since the original lexicon was built. The AKT++ lexicon used by Group C represented the best lexicon we could construct exploiting all the resources to hand.

Testbed

The database the participants searched was KMi Planet News; an online newspaper featuring events at The Knowledge Media Institute[2]. The basic Planet News search site incorporates a Main News page, showing the most recent stories (see figure 1). From this the user can access News Archive pages (B), which have a reverse chronological listing of all the stories with a drop down list that allows the user to select only one category of stories at a time, and a Search option (C) which permits simple keyword searches and advanced searches in which the user can search for authors, titles, stories or all (the default) and keywords can be combined with categories.

The baseline system, used by Group A, was this KMi Planet interface with no additional features. Group B (Magpie/AKT), used the same interface augmented with the Magpie system using the original AKT lexicon with four upper level categories: Person, Project, Research-Area and Organization. This set up is shown in Figure 1 (A).

Table 1. AKT++ lexicon by category and source

Category	AKT	KMi Portal	PANKOW
Event	0	0	74
Technology	0	21	75
Place	0	0	105
Organization	154	474	237
Person	3182	633	120
Politician	0	0	23
Company	0	0	53
Project	192	74	70
Research Area	151	92	9

Group C (Magpie/AKT++) also used KMi Planet augmented with Magpie but this time with a lexicon built from various manually generated and automatically extracted sources: the AKT lexicon used by Group B, new data from the KMi knowledge bases that had not been included in the original AKT lexicon entities extracted from the news stories by PANKOW, and entities extracted from the KMi news stories by ESpotter. The most relevant manual additions for the tasks we tested are the names of new projects. There are also 633 Persons added from the KMi Portal data. However it is worth noting that, because they are derived from KMi databases, these are KMi related people and not the kind of people we were looking for in Task A, who are visitors. This lexicon had nine upper level categories: Person, Project, Research-Agenda, Organization, Place, Event, Politician, Technology, Company. The various additions were merged with the AKT lexicon of 3679 items to create a cumulated lexicon of 6340 items, which we dubbed AKT++. We did not attempt to remove duplicates when merging since Magpie highlighting is very efficient & duplicates would not adversely affect performance. A breakdown of the numbers of entities in each of the categories supplied by each source for the AKT++ lexicon is given in Table 1.

Participants and Tasks

The participants were a mixture of research students (all working either in KMi itself or in the Open University Maths and Computing Department) and non-phd qualified researchers employed by KMi. Consequently, they all had web-searching skills and a reasonable knowledge of the subject domain. Group A contained six participants, Group B, seven, and Group C, seven.

Each participant was given a demonstration of the interface they would be using and was then asked to do two timed fact retrieval tasks in succession, which they completed in the presence of an observer.

In the "People" task participants were asked to compile a list of important people who have visited the institute. This task was designed to test the capabilities of both the information extraction tools. Since the hand-crafted sources were taken from KMi knowledge bases they mainly contained information on members of staff and students. The task was looking for visitors, whose names would not be expected to appear in the knowledge bases.

In the "Technology" Task they were asked to compile a list of technologies, either in-house or external, used in KMi projects. This task mainly tested the PANKOW system. ESpotter bases its lexicon for finding projects on the content of the KMi ontology. Therefore we did not expect ESpotter additions to make a significant difference to this task. The participants' answers had to come from the Planet News stories and they were allowed 10 minutes to complete each task. The participants recorded their answers by cut and pasting items from the stories into a text file. During the tasks the participants' interactions with the interface were recorded using Camtasia Studio, a screen recording package produced by TechSmith[3].

RESULTS

We present the following results from this evaluation: summary statistics from an analysis of the quantity and quality of items retrieved by each group, an analysis of how many of the items each group retrieved were in one of the two lexicons, and an analysis of interactions with the tools acquired from the Camtasia movies.

[2] http://news.kmi.open.ac.uk/kmiplanet/

[3] http://www.techsmith.com/products/studio/default.asp

Retrieval Performance

The first question we examined was whether having Magpie annotation available improved the participants' performance in terms of the number and quality of items they retrieved in the time available.

In order to evaluate the participants' performances on the two tasks we needed an independent assessment of the value of each item that was given as an answer to one of the questions. To do this two cumulated lists were produced which contained the 134 people and 133 technologies that had been identified by at least one of the participants and placed in an answer. These lists were presented to an impartial assessor, who was a long serving member of KMi and who had not been involved in the design or running of the experiments. He rated the items for the People task 0 (unimportant or unrecognised), 1 (moderately important) or 2 (important), and for the Technologies task 0 (not a technology or unrecognised), 1 (not an innovative technology) or 2 (innovative technology). The total value of scores that he applied for the 134 People was 94 whereas for the 133 Technologies he gave scores worth a total of 140.

Scores for each participant were calculated by summing the scores for all their answers. Mean scores for the three groups on both tasks are presented in Table 2. It is clear that both the groups using Magpie achieved higher scores for both tasks than the baseline group. Group B (Magpie/AKT) did best on the People Task, whereas Group C (Magpie/AKT++) did best on the Technologies task.

Table 2. Mean scores for the People and Technologies tasks

Task	Group A	Group B	Group C
People	13.2	15.3	13.7
Technologies	19.2	23.4	26.7

The differences between the scores for the People task are fairly small. None of the differences between groups are significant at the 5% level in two sample t-tests. Contrary to our expectation, Group C, who as we will see in section 4.2, had more highlighting available to them, scored an average of 1.6 less than Group B. Apart from the "Politician" class in the AKT++ lexicon, the highlighting made no distinction between "important" people and the rest. Therefore the more complete highlighting only tackled part of the task. The participants still had to make a judgment about which names to include. Since they were a mixture of students and young researchers they all had limited experience of the institution and about the same skill level on this part of the task. This may have been a factor in keeping the scores similar for the three groups.

These are of course small sample groups, so it only takes a couple of individuals with efficient search strategies to increase the average score of the whole group. For the People task one highly efficient strategy exploits the fact that most KMi Planet stories about a VIP visit begin with a sentence like "Ms. Ruth Thompson, the newly appointed Director of the Higher Education Strategy and Implementation Group visited the OU today.". These first sentences appear in the results listings for a keyword search so that several names could be harvested quickly by a tactical searcher. For the People task it seemed that the skill of individual searchers had a stronger effect on their final scores than whether or not they had a particular lexicon available.

The results for the Technology task, in which there was no such stylistic pattern to lead the participants to the part of a story where an answer could be found, are much more clear cut. Having Magpie annotations available increased the scores of both Groups B and C compared to Group A, and Group C, which had the enhanced AKT++ lexicon, had the highest score of all. For this task two sample t-tests showed that the difference in performance between Group A and Group C was significant at the 5% level.

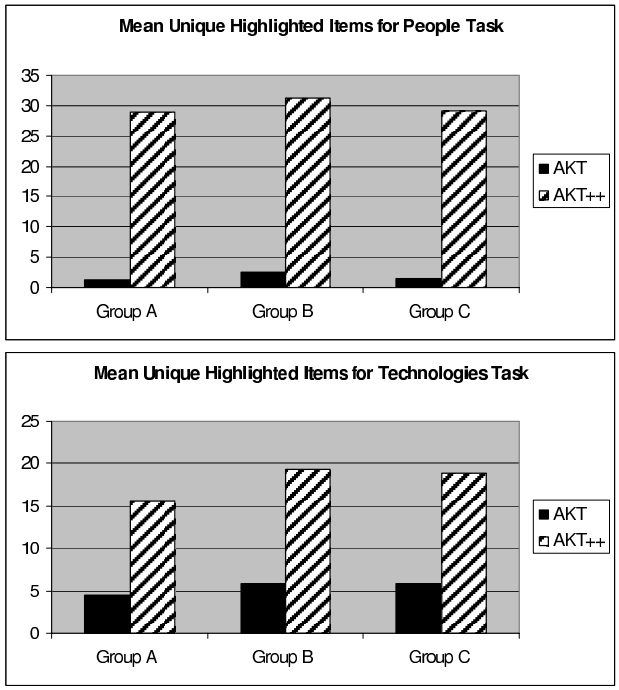

Figure 2. Number of answers covered by the lexicons for the People and Technologies Tasks

Answer Coverage

In this part of the analysis we compared how "good" the two lexicons (AKT and AKT++) were for answering the questions. To do this we determined how many of the items the participants copied into their answers were in one of the two lexicons. To do this every participant's answer was turned into a simple html file with answers emboldened. These were then viewed in a browser with each of the Magpie lexicons enabled in turn. The number of unique items highlighted for each category was counted for each task. This analysis puts the answers of all three groups together, whether they actually used the Magpie system or not, and it includes everything the participants pasted into their an-

swers irrespective of whether the items were judged to be correct. It is an assessment not of the quality of the lexicons per se, but of their overlap with the answers the participants gave. Combining the results gave us the largest possible sample.

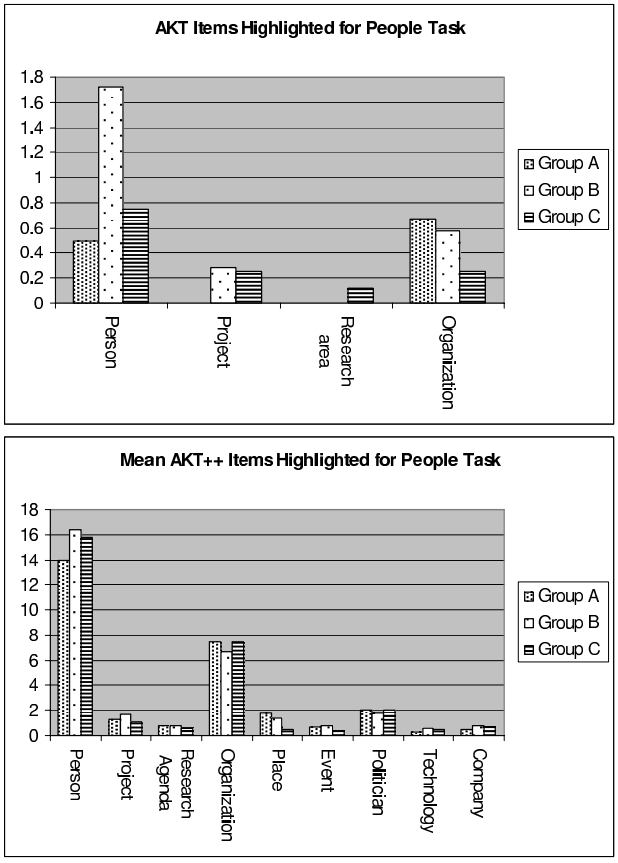

Figure 3. Breakdown by category of number of answers covered by the lexicons for the People task

For all three groups and for both tasks we found that the AKT++ lexicon highlighted more items per answer than the AKT lexicon (see figure 2). For all six cases the differences were significant at the 2.5% level in two-tailed T-tests. This indicates that the AKT++ lexicon was better suited to the tasks than the AKT lexicon; it would have given more suggestions.

Since the AKT++ lexicon has nearly twice as many entries as the AKT lexicon (6340 as compared to 3670) this result is not very surprising, but it is reassuring to know that the additional entries have potential to enhance the users' experience of the system. For the People task the difference in the answer coverage is largely due to lexicon items generated either by PANKOW or by ESpotter, since we know that the majority of names sourced from new additions to knowledge bases are those of KMi related people not visitors. For the Technology task we did a fine grained analysis which determined that 19 of the answers categorized as "Project" or "Technology" could only have been highlighted because of additions to the lexicon by PANKOW. These 19 answers scored 15 using our assessor's ratings. Typical good quality additions were "XML", "Topic Maps", "SMS" and "Semantic Web". They seem to represent technologies that are important to KMi but were "not invented here" and therefore do not appear in the institutional ontology. PANKOW is giving a qualitative improvement to the scope of the annotations, even though it is not contributing many new items.

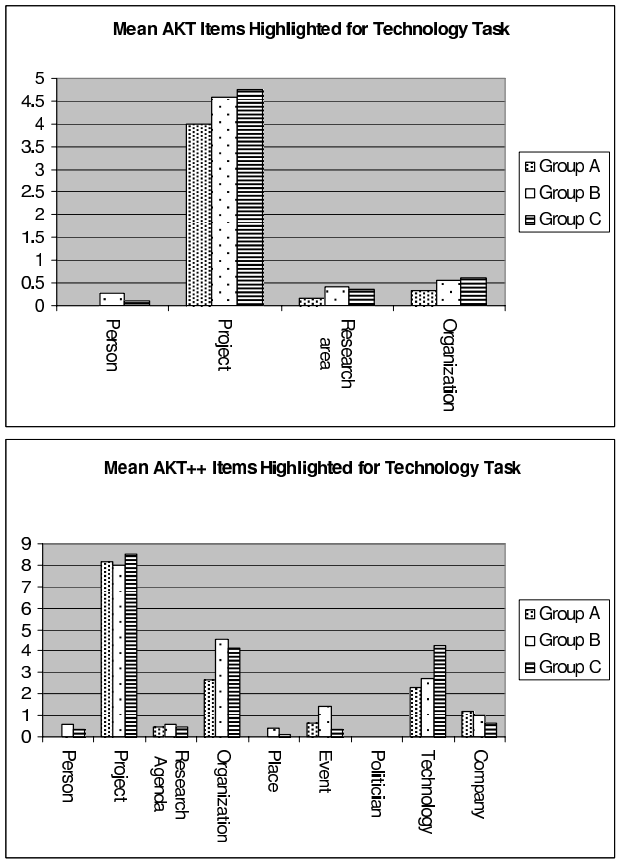

Figure 4. Breakdown by category of number of answers covered by the lexicons for the Technology task

A more interesting question was whether the highlightable entities in the AKT++ lexicon were automatically assigned to appropriate categories, which a user could reasonably be expected to choose in their search. The totals are broken down on a per category basis in Figures 3&4. The categories are the upper level categories of the lexicons which appear as buttons on the Magpie toolbar and would each be highlighted in a different colour.

The AKT++ highlighting for the People task was primarily of category Person. On average the AKT lexicon highlighted very few items for this task. The highest mean was 1.7 items highlighted in Group B's answers, this compares to 16.4 items for Group B with the AKT++ highlighting, a degree of magnitude more. This is because most of the people represented in the knowledge base work for KMi. High-

lighting these is not helpful when searching for visitors. AKT++, which was populated in part using named entity recognition methods, highlighted a much wider range of kinds of people, including visitors. The category Politician, extracted by PANKOW, which we had expected to be useful, only occurred about twice on average. There is some noise for both lexicons, with categories such as Project and Technology receiving a few hits, but the only other regularly occurring category was Organization. This appeared because the participants were given a free hand in how much text they copied into their answers. Many of them copied visitors' affiliations as well as their names. The ability to highlight affiliation is useful in the context of the People task so we consider this a positive qualitative outcome.

The AKT lexicon identified more of the retrieved items for the Technology task than it did for the People task. Means above 4 were found for all three groups for category Project. The other three categories highlighted very few items for this task. This result is reasonable since within KMi many technologies are named after the project that developed them and the projects are represented in the knowledge base used to generate the AKT lexicon. Compared to the AKT lexicon the AKT++ lexicon highlighted about twice as many items as projects with averages of 8 or above. It also highlighted noticeable numbers of Technology and Organization items. The organization names appeared in this case because many technologies include the name of the company that developed them, e.g. "Macromedia".

Overall we are satisfied that the AKT++ lexicon produced partially by information extraction methods was relating highlighted items to categories which would tend to be selected by users attempting to answer these questions. We conclude that the AKT++ lexicon is better suited to carry out the tested tasks than the AKT lexicon in terms both of coverage and categorization of items.

Movie Analysis

In Magpie the highlighting had to be refreshed for each new document that was viewed. Therefore we were able to judge how useful the participants found different highlighting options by seeing whether they used them repeatedly or whether they gave up on them after a few unfruitful trials. The Camtasia movies recorded during the experiment were analyzed to see how often the participants selected each of the Magpie highlighting options. Figure 5 presents the mean usage of the different highlighting options for Group B and Group C (Group A did not use Magpie). The most used highlighting options for Group B are Person and Project. For Group C the most used options are Person, Project, Politician and Technology.

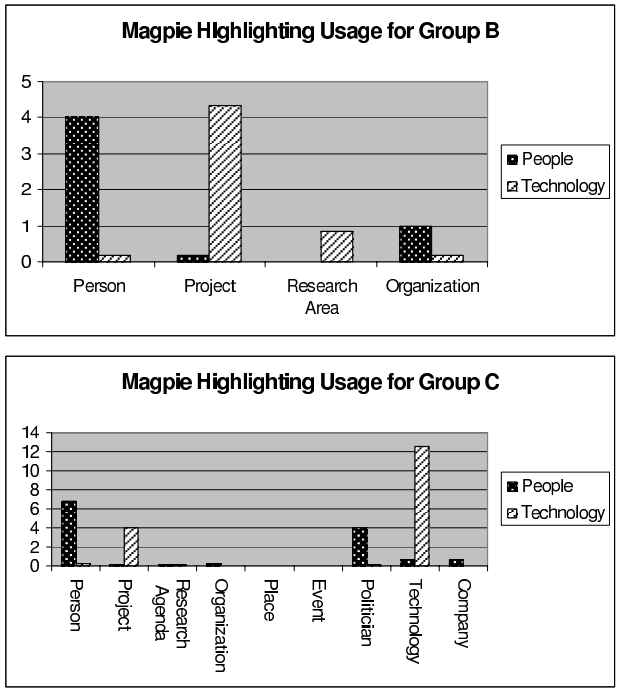

Figure 5. Magpie highlighting usage for Groups B and C

The breakdown of usage by task is similar to the answer coverage results. For the People task, Group B mainly used the Person and Organization options and Group C used Person and Politician as for the answer coverage. However, for the Technology task while Group B used mainly Project highlighting Group C used primarily Technology, even though the answer coverage indicates that using Project would have been a better strategy. We will discuss this observation further at the end of this section.

The movies were also analyzed for two additional kinds of data. We counted how often Groups B and C made a selection (typically by cut and pasting an item to their answer list) when Magpie highlighting was on, and how often, on these occasions, the item they selected had been highlighted by Magpie. This data is presented in Table 3.

The data confirms that Group C were more inclined to turn Magpie on than Group B; for both tasks the percentage of selection events that occurred with Magpie highlighting on was higher for Group C. For the People task it is very clear why. For Group B turning Magpie on gave a very low rate of return, less than one tenth of items selected with highlighting on were actually highlighted by Magpie. For Group C three quarters of people selected with highlighting turned on had been highlighted by Magpie. For the Technology task the results are very interesting because although Group C were more inclined to turn Magpie on they were actually getting a lower rate of return than Group B (43.8% c.f 65.0%). It seems that the trust built up in Group C's initial positive experience with the People task persisted into the Technology task even though the reward rate dropped.

A re-examination of the categories of highlighting people chose shows how this situation arose even though the AKT++ lexicon was a superset of the AKT lexicon, and so we would expect it to highlight at least the same number of items. Group C favored the Technology highlighting option over the Project option which actually highlighted more good selections. It is probable that the group were influenced by the match between the task requirement, to find technologies, and the label.

Table 3. Percentages of occasions when an item was selected when Magpie was switched on and percentages of occasions when highlighting was on, and the selected item was highlighted

Task	Group	% selections Magpie on	% selections highlighted if Magpie on
People	Group B	11. 7	7.1
	Group C	51.8	74.1
Technology	Group B	24.1	65.0
	Group C	62.9	43.8

CONCLUSIONS

The results gathered in this evaluation suggest that, for fact finding exercises of this kind, appropriate Magpie highlighting can help users identify more, good items in a fixed time. Comparing the two different lexicons AKT and AKT++, people are more inclined to use the highlighting for AKT++ which had been boosted with items extracted from text. With more terms highlighted they had more trust in its ability to help them and were more inclined to carry on using it. We conclude that, for the Magpie system, liberal semantic annotation, including the slightly noisy kind inevitably produced by IE, seems to work better than small amounts of high quality, human generated annotation with limited domain scope. While we cannot generalize too far beyond the scenario investigated here, our results support the case, put forward by authors such as Dill, Popov and Ciravegna [6][11][3], that IE has potential to help overcome the knowledge acquisition annotation bottleneck faced by the semantic web.

As this paper presents what, to our knowledge, is a novel method for evaluating Semantic Web applications, which we believe is complementary to existing methods for measuring the performance of information extraction algorithms, we need to reflect on our methodology as well as the results.

The core of our method was to examine the fitness for purpose of the Magpie lexicons via a task-based user study. This general philosophy should be transferable to other kinds of semantic web applications. In this experiment we looked at a search task and were therefore able to use information retrieval criteria to assess outcomes. In particular we designed test tasks which could be assessed in terms of the completeness of answers. This is only one of a number of criteria which could be applied to semantic web tasks, although it is probably the simplest to measure. Demonstrating that a system can help users to find the best answer that matches a set of complex criteria or the best way to orchestrate web services to achieve a particular goal are greater challenges that will require more sophisticated evaluation methods.

The second key point was to judge the validity of the participants' selections by post hoc examination by an independent assessor. While the ideal would be to establish a gold standard of correct answers before the experiment there is precedent for our approach in the IR community, e.g. in the TREC 2002 Arabic/English CLIR track [10].

With hindsight, the experiment could have been improved by more strict control of the content of lexicons. In particular, the addition of manual entries as well as the ESpotter entries in the KMi Portal data made these results difficult to analyze. If the groups had looked at lexicons enhanced only by either PANKOW or ESpotter annotations a performance comparison of the two IE systems would have been possible.

In terms of the time required, both the retrieval performance and answer coverage methods gave useful results with moderate effort. For a larger study both methods could be automated by writing scripts to analyze the participants' answers. The Camtasia movies also gave useful information, but analyzing them was very time-consuming. The Human Computer Interaction community has developed keystroke logging methods to record participants' actions in usability studies and these would be worth investigating for future studies.

RELATED WORK

In this section we briefly review other work concerned with exploiting metadata and annotations in order to enhance information search. In fact, we feel that currently the exploitation of metadata for information retrieval purposes is still in its infancy. Kogut and Holmes [12], for example, hypothesize that formal annotations with respect to an ontology can have an impact on information retrieval (IR), simple questions answering (Q&A) as well as for complex question answering. However, they do not present any concrete results on a task to corroborate this hypothesis. Welty and Ide [13] present a more concrete approach in which metadata is seen as instances of certain DL concepts (the authors use CLASSIC as DL formalism) and thus can be retrieved relying on DL subsumption. This makes it possible to answer queries such as 'Find all letters sent to people in Philadelphia', i.e. asking for all the instances subsuming the concept of a letter whose Recipient is in Philadelphia. However, they also do not present any concrete results on a task. In a recent survey paper [8], Hearst has argued that metadata can be used to structure information in order to enable a user to iteratively narrow down the search space in an effective and efficient way. The only studies known to

the authors in which the impact of using metadata on retrieval performance is measured experimentally with user studies are the ones in [4] and [5].

Deniman et al. [4] present a case study similar to ours in which people were confronted with an imaginary task in which they had to prepare for courses and retrieve learning material. The main measure considered by the authors is the number of actions required by the users to retrieve a relevant document. They conclude that a system exploiting metadata and text at the same time produces less deviations in the number of actions required thus being more consistent. Denoue and Vignollet [5] conducted an experiment with users, demonstrating that metadata help people in structuring their bookmarks. In a further experiment they also show that using metadata about documents also leads to improved automatic classification of documents compared to using only the text itself. Concluding, it is certainly valid to claim that research on using metadata to enhance information retrieval is still in its infancy and that, besides initial blueprints, not much work goes further than hypothesizing about the impact of metadata for search tasks. In this line our research has contributed a further step towards understanding the potential of metadata to help people in retrieving knowledge.

ACKNOWLEDGMENTS

This work was funded by the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), and the Designing Information Extraction for KM (Dot.Kom) project. AKT is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. Dot.Kom is sponsored by the European Commission as part of the Information Society Technologies (IST) programme under grant number IST-2001034038.

We also thank all the participants in the experiment, as well as Arthur Stutt for providing the impartial assessment of the retrieved items, Jianhan Zhu for ESpotter support and Laura Goebes for her help with analyzing the movies

REFERENCES

[1] Cimiano P., Handschuh S., Staab S. (2004) *Towards the self-annotating web*. In Proceedings 13th International World Wide Web Conference, (WWW 2004), May 17-22, 2004, New York, NY.

[2] Cimiano P., Ladwig G., Staab S. (2005) *Gimme' The Context: Automatic Semantic Annotation with C-PANKOW*, In Proceedings of the 14th International World Wide Web Conference, (WWW 2005), May 2005, Chiba, Japan.

[3] Ciravegna F., Chapman S., Dingli A., Wilks, Y. (2004) *Learning to Harvest Information for the Semantic Web*, In Proceedings 1st European Semantic Web Symposium, Heraklion, Greece, May 10-12, 2004

[4] Deniman D., Sumner T., Davis L, Bhushan S., Fox J. (2003) *Merging Metadata and Content-Based Retrieval*, In. Journal of Digital Information 4(3).

[5] Denoue L., Vignollet L. *Personal Information Organization using Web Annotation*, In the *WebNet 2001 World Conference on the WWW and Internet,* 2001

[6] Dill S., Eiron N., Gibson D., Daniel D., Guha R., Jhingran A., Kanungo T., Rajagopalan S., Tomkins A., Tomlin J.A., Zien J.Y. (2003) *SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation*, In Proceedings 12th International World Wide Web Conference (WWW 2003), Budapest, Hungary, 20-24 May, 2003.

[7] Dzbor, M. - Domingue, J. B. - Motta, E.: *Magpie - towards a semantic web browser*, In Proc. of the 2nd Intl. Semantic Web Conf., October 2003, Florida US

[8] Hearst M. *Next Generation Web Search: Setting our Sites*, In IEEE Data Engineering Bulletin, Special Issue on Next Generation Web Search, 2000.

[9] Kogut P., Holmes W., *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages*, In KCAP'01 Workshop on Knowledge Markup and Semantic Annotation, 2001.

[10] Oard, D.W., Gey, F.C. *The TREC-2002 Arabic/English CLIR Track*. in Eleventh Text RE-trieval Conference (TREC 2002). 2002

[11] Popov B., Kiryakov A., Ognyanoff D., Manov D., Kirilov A., Goranov M. (2003) *Towards Semantic Web Information Extraction*, In Human Language Technologies Workshop. At 2nd International Se-mantic Web Conference (ISWC2003), 20 October 2003, Florida, USA.

[12] Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F. (2003) *MnM: A Tool for Automatic Support on Semantic Markup*, KMi Technical Report, TR Number133, Sept. 2003.

[13] Welty C., Ide N. *Using the right tools: enhancing retrieval from marked-up documents*. J Computers and the Humanities 33(10):59-84, 1999.

[14] Jianhan Zhu, Victoria Uren, and Enrico Motta. *ESpotter: Adaptive Named Entity Recognition for Web Browsing*. Proc. of Workshop on IT Tools for Knowledge Management Systems at WM2005 Conference, Kaiserslautern, Germany, April 11-13, 2005.