Natural Language Interface Put in Perspective: Interaction of Search Method and Task Complexity

QianYing Wang, Jiang Hu and Clifford Nass

Department of Communication, Stanford University,
Stanford, California, USA

Abstract. A 2x2 mixed design experiment (N=52) was conducted to examine the effects of search method and task complexity on users' information-seeking performance and affective experience in an e-commerce context. The former factor had two within-participants conditions: keyword (KW) vs. natural language (NL) search; the latter factor had two between-participants conditions: simple vs. complex tasks. The results show that participants in the complex task condition were more successful when they used KW search than NL search. They thought the tasks were less difficult and reported more enjoyment and confidence with KW search. In the meantime, simple task participants performed better when they used NL rather than KW search. They also perceived the tasks as easier and more enjoyable, and had higher levels of confidence with the results, when NL was used. The findings suggest that NL search is not the panacea for all information retrieval tasks, depending on the complexity of task. Implications for interface design and directions for future research are discussed.

1 Introduction

From punch cards to keyboards to graphical user interface (GUI) to voice interface (VUI), computer interfaces have evolved to allow increasingly intuitive and natural interactions between users and computers. Among all the breakthroughs and improvements, the use of natural language (NL) as a means of input and output during human-computer interaction (HCI) is one of the most researched areas. The promising future of NL-based conversational interfaces (especially with the presence of computer agents) has been widely lauded by visionaries such as Brenda Laurel [1]. Although no one has yet to be able to claim complete success in natural language generation and processing, progress is being made every day: from text-based software agent (e.g., Microsoft™ Clippy) to speech-recognition customer services automation (e.g., United Airlines' flight information hotline), NL-based technologies have advanced into our daily life.

With the explosion of information brought by the Internet and computers in different forms, information retrieval has become a hot issue that concerns researchers and general users alike. Along with the phenomenal increase of data storage capability, information retrieval interfaces have become a research focus that is gaining more

Wang Q., Hu J. and Nass C. (2005).
Natural Language Interface Put in Perspective: Interaction of Search Method and Task Complexity.
In *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science*, pages 3-12
DOI: 10.5220/0002566000030012

attention than ever before. In the HCI research community, the literature about seeking and interacting with information is incredibly rich with design principles and maxims [2-11]. However, most of those usability research projects were focused on keyword (KW) search. In the meantime, NLP researchers have focused on building robust knowledge based system [12-16]; little attention has been given to interface issues.

Natural language search has been made possible in some specialized areas such as U.S. legal materials [17] as well as in everyday contexts such as askjeeves.com [18]. Users may type full sentences when natural language search is available. Boolean search, on the other hand, has thus far been the dominant means of information retrieval. However, our observation and informal interviews reveal that average users are not familiar with Boolean search. Most of the time what they use are just keywords or key phrases without operators such as "+" and "/". For an average user, "Boolean search" quite possibly equals to "keyword search."

The most relevant piece of literature to the present paper is Turtle's [19] study comparing natural language and Boolean query. In two experiments comparing information retrieval performance from collections of full-text legal materials, Turtle found that searchers using a then current-generation natural language system were more successful than expert users who relied on Boolean search. The results seem to herald the proliferation of natural language search systems.

However, there is one major problem with Turtle's studies: the natural language queries were not free-will sentences made up by the searchers. Instead, all "natural language" queries were guided by a set of issue statements made available to the searchers. It is quite likely that these "trained" queries boomed the performance of natural language understanding. As a result, the conclusion of Turtle's study has to be interpreted carefully.

In this paper, we attempt to compare natural language and keyword search within a more everyday context: e-commerce customer support, with untrained average users. We are also interested in how task complexity affects users' interaction with information retrieval interface. Our hypothesis is that *increased task complexity will significantly undercut users' performance with natural language system*, because it is difficult for users to turn complicated search requests into concise NL queries that can be understood by the computer.

2 Method

2.1 The Two Interfaces

The natural language search interface was the AIM chat-room, with a natural language agent created by Conversagent, Inc. [2]. Popular search interfaces such as askjeeves.com encourages users to type keywords or questions on the same webpage. However, we believe that online chat interfaces such as that of AIM is much more conducive to natural language input than is a small text box on a webpage. Agents available for AIM users included My eBay Buddy, AgentBaseball, EllegirlBuddy and SmarterChild. My eBay Buddy answers frequently asked questions about buying and

selling on eBay. AIM users can add these agents to their buddy lists so as to chat with them and acquire information from them.

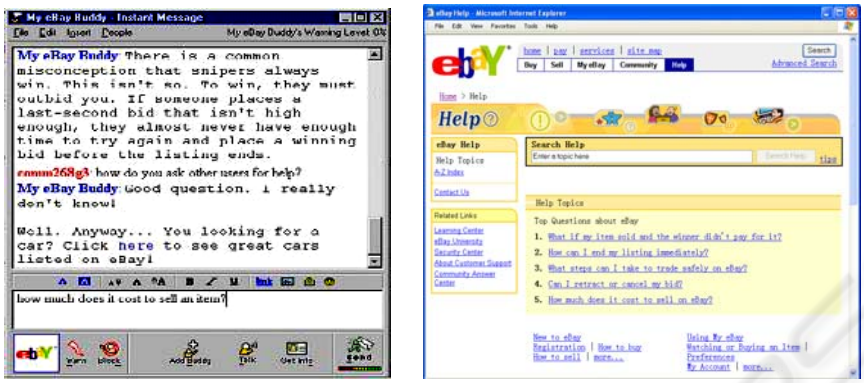

Fig. 1. Screenshots of AIM chat window with My eBay Buddy, and the eBay Help website

For the keyword search condition, we used the eBay Help website, the parallel website to My eBay Buddy. Figure 1 shows the two interfaces side by side. Search results provided by My eBay Buddy were presented in a conversational manner while eBay Help website returned results in a typical list of ranked links. We are fully aware of the differences between the two presentation formats. However, these forms of presentation are essential parts of the two radically different interfaces and should not be studied separately. Behind the interfaces, the eBay Help website and My eBay Buddy shared the same database. That is to say, two identical queries with the two interfaces would get the same information.

2.2 Search Tasks

There were two task categories: complex and simple. Complex search tasks included eight questions concerning eBay transactions; the number of simple search tasks was 20. Compared to simple tasks, complex tasks were more difficult with either NL or KW interface. Finishing a complex task often involved integrating information from different aspects related to the task. For example, one complex task was to find out how much people must pay eBay if they were selling a $35 merchandise. Users had to find information about listing fee, final value fee, eBay picture service fee, etc., and combine the information for a correct answer. Our pre-test indicate that people spend about the same amount of time to finish eight complex tasks as to finish twenty simple tasks.

The following are two of the simple tasks used in the present experiment:

- *Please find out what the gift icon means.*
- *Please find out what the different colored stars mean.*

In the meantime, complex tasks look like these:

- *Please find out if it is legal to place a bid right before the auction closes.*

- *Please find out what you should do if you don't want a certain person to bid on your item.*

For a full list of search tasks used in the present study, please visit http://www.stanford.edu/~wangqy/projects/tasks.htm.

2.3 Participants

All native speakers of English, 52 college students (N=52) from an introductory communication class participated in the experiment for course credit. They were randomly assigned to all conditions, with gender balanced across conditions. None of the participants had ever sold anything or bought more than three items on eBay.com. Experiences with AIM were also balanced across conditions. All participants signed informed consent forms upon arrival at the lab and were debriefed upon the completion of the experiment.

2.4 Procedure

The experiment had a 2x2 mixed design, with search complexity (simple vs. complex) as the between-participants factor and search interface (keyword vs. natural language search) and the within-participants factor. Participants performed search tasks in a research laboratory equipped with personal computers. Upon arrival to the laboratory, each participant was seated and assigned to a computer with both Internet Explorer and AIM.

After reading instructions on–line, participants were given a list of either simple or complex tasks about buying or selling items on eBay. For half of the tasks, participants used keyword search on the eBay Help website (i.e., KW); for the other half, participants conversed with My eBay Buddy (i.e., NL) to find the answers. Half of the participants started with KW and the other half started with NL. Participants typed in an answer when they thought they had found one, indicating whether they believed that they had found the correct one. Upon finishing each task, participants responded to several questions about the finished task.

2.5 Measures

Actual performance was computed as the percentage of questions that the user answered correctly. *Perceived performance* was computed as the percentage of questions that the users believed that they had answered correctly. *Time on task* was the average amount of time completing each search.

Questions concerning perceived task difficulty, pleasure, and confidence in working with the search interface were asked for each task. Participants used radio buttons to indicate their responses for these questions. Each question had an independent, 10-point Likert scale. *Perceived task difficulty* was an average across searches of responses to the questionnaire item, "How difficult did you feel the search was?" *Enjoyment* was an average across search questions of responses to the questionnaire

item, “How enjoyable did you find the search?” *Confidence* was an average across search questions of responses to the questionnaire item, “How confident were you with your answer?

3 Results

3.1 Manipulation Check

Time to finish complex tasks and simple tasks were recorded during the study. Consistent with our manipulation, complex searches took longer than simple search for both keyword [$F(1, 24)=107.6$, $p<.001$] and natural language [$F(1,24)=55.6$, $p<.001$; see Table 1]. There was no statistical difference for *time on task* between the keyword search and natural language search ($p>.05$).

Table 1. Time on task

(minutes)	Using NL	Using KW
Simple	1.21	1.33
Complex	3.8	4.17

3.2 Actual vs. Perceived Search Performance

Figure 2 shows results for actual and perceived performance. There was a significant interaction between task complexity and search interface for both perceived [$F(1,50)=31.4$, $p<.001$] and actual performance [$F(1,50)=26.5$, $p<.001$]. Complex task participants were more successful in finding the needed information with KW than with NL; they also perceived more success in the KW condition. Conversely, simple task participants were more successful with NL than with KW, and they perceived more success in the NL condition.

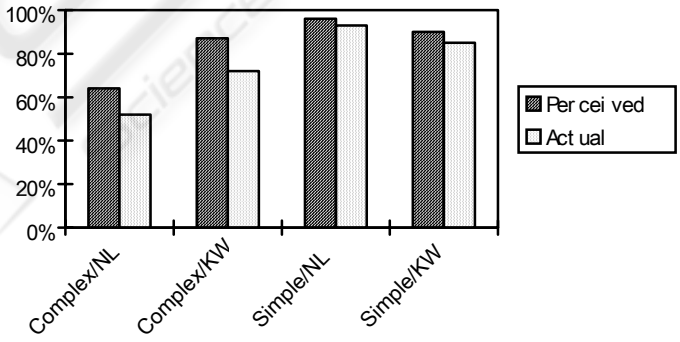

Fig. 2. Accuracy of search results

Simple task participants obtained [M_{simple}=89%, $M_{complex}$=62%, F(1,50)=50.1, p<.01] and believed they obtained [M_{simple}=93%, $M_{complex}$=76%, F(1,50)=41.6, p<.01] more correct answers than complex task participants. Not surprisingly, there was a tendency for participants to over-estimate their actual performance. Though there was no significant difference between the two search methods for actual success [M_{KW}=79%, M_{NL}=73%, F(1,51)=3.24, p>.05], participants thought they obtained more correct answers from keyword search than from natural language queries [M_{KW}=80%, M_{NL}=89%, F(1,51)=7.61, p<.01].

3.3 Perception of Task Difficulty, Task Enjoyment, and Confidence with Answers

Perceived Task Difficulty. We first analyzed participants' perception of difficulty without including performance results as covariates. There was a significant interaction effect between task complexity and search interface on participants' perception of difficulty [F(1,50)=105.0, p<.001]. Complex task participants thought KW search was easier to work with than was NL; conversely, simple task participants thought NL was easier to work with than was KW search (see Figure 3). High complexity tasks were perceived to be more difficult than were low complexity tasks (M_{simple}=3.20, $M_{complex}$=4.95, F(1,50)=36.4, p<.001). No main effect was found for search interface [F(1,50)=.917, p>.34].

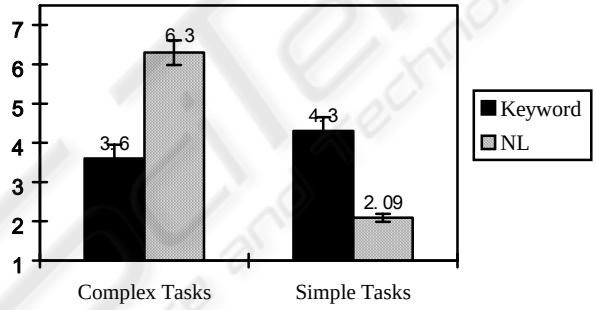

Fig. 3. Perception of difficulty

We re-examined the *perceived difficulty* data with *actual performance* as covariates. The analysis reaffirmed our finding of an interaction effect between task complexity and search interface on perceived difficulty [F(1,50)=28.95, p<.001]. That is, even after adjusting the actual performance score, keyword search was preferred for complex tasks and natural language was preferred for simple tasks.

Task Enjoyment and Confidence with Answers. Participants' perception of enjoyment and confidence with the search interfaces were analyzed. Table 2 presents the means.

Table 2. Perception of enjoyment and confidence

	Complex Task		Simple Task	
	NL	KW	NL	KW
Enjoyment	4.07	5.43	6.9	4.84
Confidence	5.84	8.03	8.71	7.65

There was an interaction effect between interface and task complexity with respect to task enjoyment [$F(1,50)=47.6$, $p<.001$]. Participants found the tasks to be least enjoyable while working on complex tasks using NL and most enjoyable while working on simples tasks using NL. The effect remained after controlling for actual performance [$F(1,50)=19.9$, $p<.001$]. No main effect was found for search interface on participants' enjoyment [$F(1,50)=2.01$, $p>.05$]. Overall, complex task participants found the tasks to be less pleasant than simple task participants [$M_{simple}=5.87$, $M_{complex}=4.75$, $F(1,50)=5.61$, $p<0.05$].

Consistent with participants' perception of enjoyment, there was an interaction for confidence with answers (even after controlling for actual performance): Participants were least confident while working with complex tasks using NL and most confident while working with simple tasks using NL [$F(1,50)=70.9$, $p<.001$; control: $F(1,50)=13.1$, $p<.001$].

Participant were more confident while working with KW search than with NL queries [$F(1,50)=8.56$, $p<.01$]. Interviews with participants during debrief sessions concerning confidence are discussed in the following section. Not surprisingly, complex tasks made user feel less confident than simple tasks [$F(1,50)=15.4$, $p<.001$].

4 Discussion

Results of the present study suggest that the selection of search interface has important and systematic effects on user performance and perceptions. However, these effects are conditioned by the complexity of search task and the nature of queries. The general pattern we found was that NL was better (in both reality and perception) than KW search for simple search tasks, while KW search was better (in both reality and perception) than NL for complex search tasks. The type of search interface had no direct effect on either perceived task difficulty or enjoyment, although participants were more confident with KW than with NL search. During debriefing, several participants said that they felt more confident with KW because they had browsed pages or links that gave them peripheral affirmation. It is important to note that although there was no difference in terms of actual search performance, participants believed that they obtained more correct answers by using the KW interface than the NL interface.

4.1 Task Complexity Effect and Implications

One clear problem with NL search is that even good NL systems have problems understanding users. This was true for almost all dialogs between complex task participants and the Buddy. When the understanding failed, My eBay Buddy revealed its inability to understand the participant by asking questions back to the participant, thus giving the participant an impression that he/she was engaged in a conventional conversation with the Buddy as if the Buddy was another person.

In a conventional (i.e., human-human) conversation, people adjust their communication strategies by asking and responding to questions in order to establish some "common ground" so as to enhance understanding [21]. When a computer agent tries to talk like a real person, the user would subsequently raise his/her expectations of the agent's ability to understand. Some previous research [22, 23] has suggested that users may say things more freely when they have high expectations of the agent than when they have low expectations. The freer utterances from the user consequently impose extra burden on natural language processing system. The gap between unconstrained user utterances and the limited understanding capability of NL agent thus frustrates the user and lowers user satisfaction.

When using the Buddy in the present study, participants with complex tasks in general found it difficult to form concise questions. As a result, they had to ask more questions than did simple task participants. Given the limited natural language processing capability of the Buddy, the more questions participants asked, the more likely the Buddy would fail to understand. Also, the more unconstrained the questions were, the more likely the Buddy would appear ignorant. It would be a great idea if the agent is able to constrain user utterances by leveraging the "alignment" phenomenon observed in human-human dialogs [24].

On the other hand, under the simple task condition where the NL agent was able to understand and to offer relevant responses, user satisfaction was high. In fact, during our debrief sections, several simple task participants mentioned that they felt the Buddy was really smart when it asked questions back to confirm its understanding of participants' queries. Under such circumstances, it seemed to be acceptable for the agent to behave like a human being.

The above findings suggest that a natural language agent should adapt its response style according to its confidence level in understanding user input. A low confidence level means that the agent should reduce sentential responses because large amounts of sentential responses may mislead the user to expect the same intelligent output from the agent as from a human conversant. The user's increased level of expectation may lead to a decrease in satisfaction with the agent if the agent performs poorly.

4.2 Future Research Directions

In the present study, participants were not given the freedom to choose between the two interfaces; failure rates on complex tasks were high for both NL and KW search. One potential direction for future research is to investigate the interchange between KW and NL interfaces. Task performance and perceptions are likely to be different when users are able to switch from one interface to the other if they think that the

initial interface is ineffective. The research question here is whether or not combining two search interfaces/methods could provide better user experience with information retrieval systems.

As noted earlier, it is important to understand how the response style of the agent influences user behaviors and attitudes. Some earlier studies have demonstrated that linguistic variations such as sentence length may affect users' input by soliciting alignment (i.e., mirroring) behaviors from them [22, 25]. However, linguistic alignment is a two-way process. How users would evaluate an aligning agent versus a non-aligning agent is still to be explored. On top of that, researchers may further determine when and how a computer agent should align with the user to achieve or improve user satisfaction.

4.3 Final Words

The design of search methodology requires an understanding of the complex interaction between technology, psychology, and context. While one could have a reasonable argument over whether KW search is better than NL-based information retrieval, or vice versa, the present study shows that the "KW vs. NL" argument is inadequate without reference to the particular task and the model of search that the user brings to the task.

References

1. Laurel, B.: Interface Agents: Metaphors with Character. In Laurel, B. (ed.), The Art of Human-Computer Interface Design. Addison-Wesley, Reading, MA (1990)
2. Blackmon, M.H., Polson, G.P., Kitajima, M, Lewis C.: Design Methods: Cognitive Walkthrough for the Web. Proceedings of CHI 2002. ACM Press, 463 – 470
3. Borgman, C.L., Belkin, N.J., Croft, W.B., Lesk, M.E., Landauer, T.K.: Retrieval Systems for the Information Seeker: Can the Role of Intermediary be Automated? Proceedings of CHI '88 (Washington, D.C.), ACM Press, 51–53
4. Card, S., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R., Schraedley, P., Boshart, J.: Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. Proceedings of CHI 2001. ACM Press, 498–505
5. Chi, E., Pirolli, P., Chen, K., Pitkow, J. E.: Using Information Scent to Model User Needs and Actions on the Web. Proceedings of CHI 2001. ACM Press, 490–497
6. Larson, K., Czerwinski, M.: Web Page Design: Implications of Memory, Structure and Scent for Information Retrieval. Proceedings of CHI '98. ACM Press, 25–32
7. Nielsen, J.: Designing Web Usability. New Riders, Indianapolis, IN (1999)
8. Pirolli, P.: Computational Models of Information Scent-Following in a Very Large Browsable Text Collection. Proceedings of CHI '97. ACM Press, 3–10
9. Pirolli, P., Card, S. K.: Information Foraging. Psychological Review 106(4) (1999), 643–675
10. Sellen A.J., Murphy R., Shaw K.L.: How Knowledge Workers Use the Web. Proceedings of CHI 2002. ACM Press, 227–234
11. Spool, J.M., Scanlon, T., Schroeder, W., Snyder, C., DeAngelo, T.: Web Site Usability. Morgan Kaufman, San Francisco, CA (1999)

12. Doszkocs T.: Natural Language Processing in Intelligent Information Retrieval. Proceedings of the 1985 ACM Annual Conference on The Range of Computing: Mid-80's Perspective. ACM Press, 356–359
13. Guglielmo E.J., Rowe N.C.: Natural-Language Retrieval of Images Based on Descriptive Captions. ACM Transactions on Information Systems (TOIS), July 1996, Vol. 14, Issue 3, 237–267
14. Jacob, P. S., Rau, L.F.: Natural Language Techniques for Intelligent Information Retrieval. Proceedings of the Eleventh International Conference on Research & development in Information Retrieval (1988), 85–99
15. Meng, F.: A Natural Language Interface for Information Retrieval from Forms on the World Wide Web. Proceeding of the 20th International Conference on Information Systems (1999), 540–545
16. Vivacqua A., Lieberman H.: Agents to Assist in Finding Help. Proceedings of CHI 2000. ACM Press, 65–72
17. Griffith, C.: WESTLAW's WIN: Not Only Natural, But New. Information Today, 9-11, October 1992.
18. Ask Jeeves, Inc. http://www.ask.com
19. Turtle H.: Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (1994), 212–220
20. Conversagent, Inc. http://www.conversagent.com
21. Clark, H.H.: Using Language. Cambridge University Press, New York (1996)
22. Brennan, S.: Conversation with and through Computers. User Modeling and User-Adapted Interaction, 1 (1991), 67–86
23. Pearson, J., Pickering, M.J. Branigan, H.P., McLean, J.F., Nass, C.I., Hu, J.: The Influence of Beliefs about an Interlocutor on Lexical and Syntactic Alignment: Evidence from Human-Computer Dialogues. Proceeding of the 10th Annual Architectures and Mechanisms of Language Processing Conference (2004)
24. Nass, C.I., Hu, J., Pearson, J., Pickering, M.J., Branigan, H.P.: Linguistic Alignment in HCI vs. CMC: Do Users Mimic Conversation Partners' Grammar and Word Choices? Unpublished manuscript (2004).
25. Zoltan-Ford, E.: How to Get People to Say and Type What Computers Can Understand. International Journal of Man-Machine Studies, 34(4) (1991), 527-547