Windows as a Second Language: An Overview of the Jargon Project

Tim Paek

Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
timpaek@microsoft.com

Raman Chandrasekar

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
ramanc@microsoft.com

Abstract

When novice users request help from expert sources, they often have difficulty articulating the nature of their problems due to their unfamiliarity with the appropriate technical terms or "jargon." Consequently, in searching for help documents, they struggle to identify query terms that are likely to retrieve the relevant documents. There is often a vocabulary mismatch between the technical language of help documents or experts and the common language of novice users. At Microsoft, as part of the Jargon Project, we are exploring methods to characterize the use of jargon by different communities of users, and techniques to exploit user models based on jargon-related features to provide customized and more appropriate help and support. This paper provides an overview of the Project and presents preliminary results demonstrating the feasibility of distinguishing types of users by their jargon use.

1 Introduction

When novice users request help from expert sources, they often have difficulty articulating the nature of their problems. This is in part due to their unfamiliarity with the appropriate technical terms, or "*jargon*," to describe the constructs and relations of their situation as well as the proper usage for those terms as utilized in a community of experts. For example, consider the following newsgroup posting: "*How do I run the Internet on Windows XP?*" Experts examining the question would surmise that the person was most likely a novice user with respect to the Internet by both the nature of the topic, which is accessing the Internet, and the atypical use of the technical term "*Internet*," in that "*Internet*" is not commonly used in a direct object relation for the verb "*run*." In other words, the question does not conform to the canonical usage, and as such, reveals that the user is most likely not a member of the "language community" of people who know enough about the topic to use that term in a "conventional" or a socially agreed upon manner (Clark, 1996).

In information retrieval, researchers have noted that a vocabulary mismatch often prevails in failed searches between the query expression and the language of document authors (Furnas et al., 1987; Deerwester et al., 1990). This mismatch is particularly evident in the retrieval of help documents, which are written in precise technical terms, and novice queries, which often employ general terms for concepts users may not fully understood. At Microsoft, as part of the Jargon Project, we are exploring methods to characterize the use of jargon by different communities of users, and techniques to exploit user models based on jargon-related features to provide customized and more appropriate help and support. In this paper, we provide an overview of the Project, describing the motivation, objectives, and tools we are using to characterize jargon usage. We also present

preliminary findings that demonstrate the feasibility of distinguishing types of users, corresponding approximately to their level of expertise, by their use of jargon.

2 Jargon Project

The Jargon Project is aimed at understanding how jargon usage affects the retrieval of help documents by users at different levels of expertise. The ultimate goal of the Project is to facilitate more personalized retrieval of help documents based on user models inferred from a history of natural language queries. The user models are geared towards inferring level of expertise in various domain topics by examining the linguistic structures pertaining to the use of specific jargon terms. We now describe how the jargon terms were selected and a canonical corpus created to help build such user models.

2.1 Data Collection

In order to create baselines for comparison, we needed to establish a canonical corpus of expert language to which natural language queries could be evaluated. The corpus collected consisted of the text extracted from all help documents written for Windows XP Help and Support Center (HSC) and Office XP products (Word, PowerPoint, Excel, Outlook). The text exceeded 100 MB and covered over 9800 individual help documents. In addition, we extracted text from the body chapters of 21 Microsoft Press books relating to the operation of Windows and Office XP. This text exceeded 23 MB and included over 500 chapters.

The canon, which constitutes a "gold standard" set of jargon terms, was compiled from the corpus using the following definition. Jargon comprises any term that either:

1) Appears in a glossary (e.g., Windows HSC glossary, or the glossary of any of the Microsoft Press books), or
2) Appears in an index of a technical publication.

The general intuition behind this definition is that if a technical term is important enough that authors of technical publications feel necessary to identify it in the index or glossary (so that users may find all references to it or its definition), then that term is considered jargon. Note that for the index, we used only the first-level headings, mostly to limit the number of jargon terms. Overall, we collected roughly 11,000 glossary and index terms, though many of these terms were redundant due to the use of abbreviation or plurals. Observing that some of the terms were abbreviated, we also parsed them to extract any parenthetical acronyms, such as DOS or ODBC. We amassed about 700 acronyms in all.

To obtain natural language queries, we collected over 1.5 GB of postings to 20 "microsoft.public" newsgroups pertaining to Windows and Office XP help and support. We selected particular newsgroups that were known to be geared towards novice users, as opposed to just expert users. In these newsgroups, most of the replies to queries typically come from Microsoft employees monitoring the postings, or MVPs (Most Valued Professionals), non-Microsoft employees officially recognized by the company as experts who have significantly contributed to helping others resolve their software problems. MVPs typically host their own websites with Frequently Asked Questions (FAQ) about particular Microsoft software, and often refer novice users to canned responses they have on their websites.

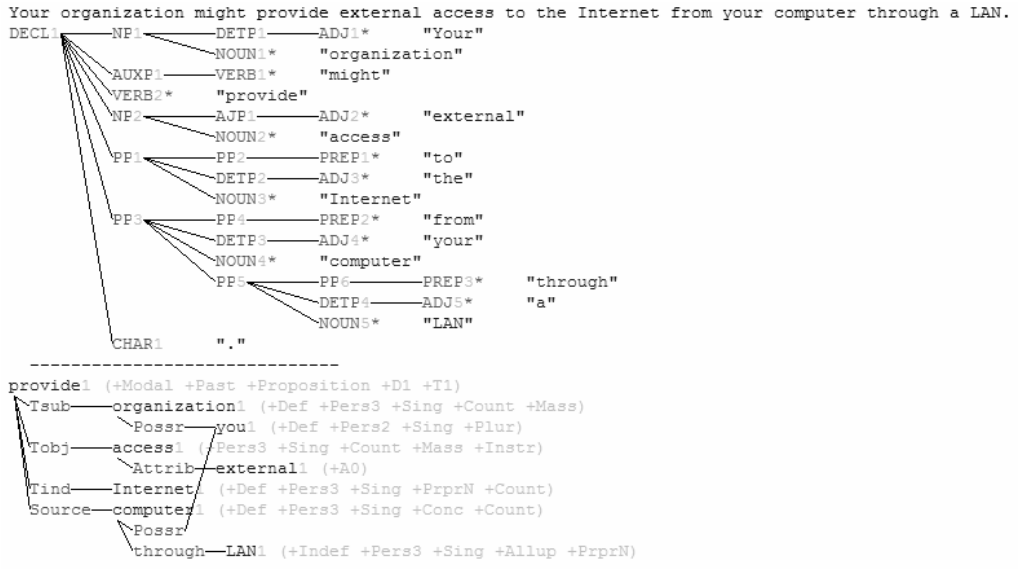

Figure 1: Parse tree and logical form for a sentence containing "Internet" in a help document.

2.2 Characterizing User Groups

Given our interest in learning user models of expertise level based on jargon features, we needed to identify groups of users that corresponded roughly to level of expertise. We divided postings from the user population into the following contrastive groups:

- Experts vs. Non-experts: where Experts were defined as users whose emails ended in either "microsoft.com" or "mvp.org"
- First-in-thread vs. Not-first-in-thread: where First-in-thread contained only postings that started off a thread of replies.
- Queries vs. Solutions: where Queries represented postings that were heuristically selected by key phrases such as "how do you," and Solutions represented replies selected by key phrases such as "Have you tried," that often pointed to known solutions in KB articles and MVP FAQ lists.

We selected the above groups since in reading the postings to the newsgroups we selected, we found that novice users tended to start off query threads, and expert users tended to respond to them. In comparing jargon usage across different types of users, we considered combinations of these contrastive groups, as discussed in section 3. Note that in extracting the content of postings, only new content was considered; that is, we removed all inclusions and indirect quotations, demarcated typically by the line prefix ">".

2.3 Generating Natural Language Features

In order to generate features relating to the usage of jargon, we first obtained natural language parses of all the sentences in the canonical corpus of expert language as well as those in the newsgroup corpus. The parser we used was Microsoft NLPWin (Heidorn, 1999), which not only

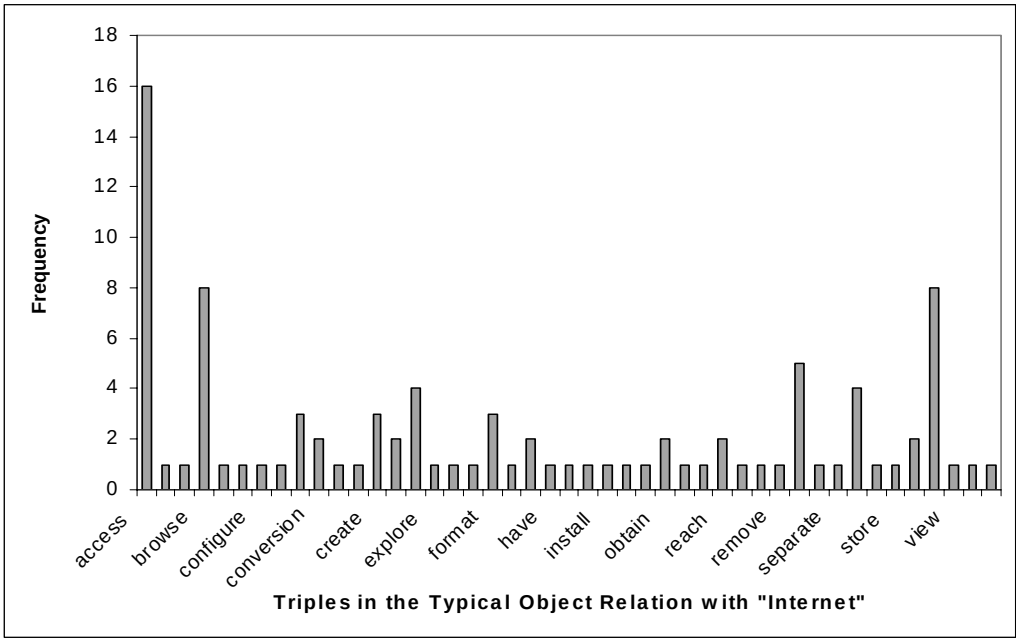

Figure 2: Distribution of words in the typical object relation with "*Internet*"

generates a syntactic tree but also constructs a logical form of the sentence from the tree, representing predicate-argument relations in a semantic graph. Figure 1 displays the parse tree and logical form for a sentence in a Windows XP help document containing the jargon term "*Internet*" in the typical indirect object role for the head "provide." Once the logical form is generated, we extract triples of the structure <*object1*, ***relation***, *object2*>, where the relations are represented labels on the arcs in the logical form, as shown in the bottom half of Figure 1 (e.g., <provide, **Tind**, Internet>, <provide, Tobj, access>, <provide, Source, computer>).

Given all triples containing a jargon term in either of the two object roles, features can be derived based on the relationship of a given token triple to a distribution of tokens. Consider again the example sentence "*How do I run the Internet on Windows XP?*" The triple for this sentence for the jargon term "*Internet*" in the typical object relation would be < run, **Tobj**, Internet>. We can generate features based on how this triple compares to a distribution of triples as found in canonical text with the same relation and second object. Figure 2 displays a distribution of triples with <*, **Tobj**, Internet> for the Windows XP HSC corpus where the word "*access*" is the most frequently occurring first object of that triple (e.g., "*access the Internet*"). In this case, we would generate a feature specifying that <run, **Tobj**, Internet> did not appear in the Windows XP HSC corpus, and hence, contained zero mass. If it did appear, we would have as our feature its mass with respect to the other triples.

3 Variation of Jargon and Function Word Use by User Community

While it may seem intuitive that users at different levels of expertise might employ jargon in contrasting ways, we sought to empirically verify the feasibility of distinguishing user expertise

level from jargon-related features. The most obvious distinction that we expect between user populations is that novice users would be much less likely to employ jargon as frequently as expert users, presumably because they are not as familiar with these terms. To verify this, we examined the contrastive groups of newsgroup postings specified in section 2.2, as well as combinations thereof, to see if any of the groups utilized more jargon terms than the rest. In this section, we present preliminary results demonstrating a clear difference in the percentage of jargon use between groups, suggesting at least one point of distinction.

	Non-Expert & First In Thread	Non-Expert & Query	Expert & Not First In Thread	Expert	Non-Expert	First In Thread	Not First In Thread	Query	Solution
jargon / words	2.12%	1.73%	4.32%	3.37%	2.21%	1.80%	1.49%	1.73%	3.57%
words / postings	84.07	117.54	99.65	133.82	103.30	105.39	114.10	117.54	91.69
jargon / postings	1.79	2.03	4.31	4.51	2.28	1.90	1.70	2.03	3.27

Table 1: Jargon usage by different types of newsgroup postings

Table 1 displays a comparison of the jargon usage between the three contrastive groups as well as three combinations of those groups. The first two columns, "Non-Expert & First-in-thread," and "Non-Expert & Query," were meant to capture what seemed to be typical behavior of novices; that is, they tend to be non-MVP, non-Microsoft employees who post a query or comment that starts a newsgroup thread. "Expert & Not First-in-thread" was meant to capture the behavior of experts, typically MVPs or Microsoft employees, who were not starting threads but answering them. To generate the jargon percentages shown in the table, we simply counted the number of times a jargon term in the canon appeared in the various groups of newsgroup postings and normalized that by the total number of words in that group. Every group had well over 2 million words. We also compared the number of words in the postings to the number of postings, as well as the number of jargon terms in the postings to the number of postings.

With respect to the contrastive groups, "Experts" use more jargon than "Non-experts" and postings heuristically defined as "Queries" used less jargon than "Solutions." The later is true despite the fact that we noticed that many solutions simply referred users to other help resources such as MVP FAQ lists and KB articles. Interestingly, the "First-in-thread" versus "Not-First-In-Thread" group did not display much of a difference; this may be due to the fact that often people who reply to a newsgroup posting simply request more information from the sender.

The key finding in these results relates to the difference between the two groups meant to capture novice behavior (i.e., "Non-Expert & First-in-thread," and "Non-Expert & Query") and the exemplary expert group (i.e., "Expert & Not First-in-thread"). The exemplary expert group was at least twice as great as the novice groups. To verify the significance of the greatest difference, we conducted a chi-square test of homogeneity between "Non-Expert & Query" and "Expert & Not First-in-thread" and found the difference to be significant (χ^2 (1, N > 14650993) = 86804.08, p < .001). While the difference in percentages may seem small, it must be examined in relation to the baseline, or the percentage of jargon terms to word tokens in the entire newsgroup corpus, as shown in Figure 3. While there is hardly any difference between the baseline and "Non-Expert &

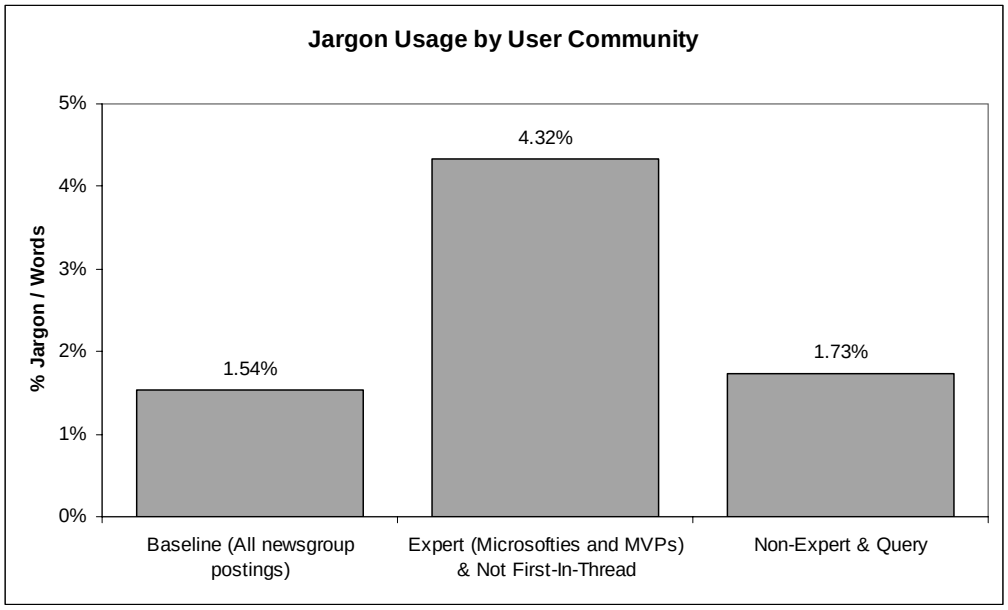

Figure 3: Incidence of jargon use by user community

Query," the difference between the baseline and "Expert & Not First-in-thread" is even greater than the aforementioned significant difference.

Aside from the percentage of jargon usage, it is interesting to note that "Non-expert & Query" postings had more words per postings than "Expert & Not First-in-thread" group. This suggests quite intuitively that while novices used more words in their queries than experts, experts used more jargon terms. Furthermore, while "Expert" postings employed more words per postings than Non-expert postings, "Solution" had fewer words than "Query," implying that experts have succinct responses to problems. One combination group which we have yet to evaluate against the baseline, is "Expert & Solution," which given the results for "Solution" and "Expert & Not First-in-Thread" is likely to have the highest percentage of jargon usage.

3.1 Function/Stop Words

Since researchers began to apply statistical techniques to infer author attribution of historical texts such as the Federalist Papers (Mostellar and Wallace, 1964), one of the most predictive stylistic discriminators (characteristics of style which remain invariant within a corpus of works for a particular author but which varies from author to author), has been the use of function words: words that serve a grammatical purpose but have no meaning by themselves (e.g., "and", "or", "of"). As features used with various machine learning techniques, function and stop words have been shown to successfully classify authors by gender (Koppel et al., 2002). So, we sought to test if function and stop words, like jargon terms, might be predictive in distinguishing between users of different level of expertise. Using the same set of 467 function and stop words from Koppel et al. (2002), we computed the percentage of function and stop words in the different types of newsgroup postings and obtained interesting results.

"Non-Expert & First-in-thread" (28.2%) differed very little from "Expert & Not First-in-thread" (29.4%), implying that function and stop words by themselves may not adequately distinguish between novices and experts; however, the postings for "Non-Expert & Query" differed significantly from both (21.4%). Here it is important to point out that all "Non-Expert & Query" postings were also "Query" postings and vice versa; there were no query postings by experts. The difference we observed in the percentage of function and stop words between "Query" (21.4%) and "Solution" postings (38.5%) is high. This difference makes sense upon a closer analysis of the distribution of function and stop words. "Solution" postings contained much more of the "you"-related function words (e.g., "you", "your") than "Query" postings. Specifically, 8.8% of all "Solution" function words were "you"-related, compared to only 1.7% of all "Query" postings. Reading the actual messages, the primary reason for this is due to the fact that in advising users to take particular actions, "Solution" postings almost always directly address the users (e.g., "you need to reboot your system").

4 Future Directions and Conclusion

In this paper, we presented an overview of the Jargon Project at Microsoft aimed at understanding how jargon usage affects the retrieval of help documents by users at different levels of expertise. We also described preliminary results demonstrating the feasibility of distinguishing types of users by their jargon usage. In particular, we found that novice users in newsgroup postings employ about half as much jargon as experts in general, presumably because they are less familiar with the proper technical terms, even though they use more words per postings. Given that the ultimate goal of the Jargon Project is to facilitate more personalized retrieval of help documents based on user models inferred from a history of natural language queries, our next step is to build classifiers using the jargon-based features, and to use the classifiers to re-rank search results for natural language queries. We also plan to compare difference classes of computer manuals, and build differential user models based on our analyses.

5 Acknowledgements

We thank Robert Ragno for his willing and patient help with text analysis algorithms. We also thank various colleagues in Microsoft and in Microsoft Press who gave us access to the help documents and the text of MS Press books.

References

Clark, H. (1996). *Using language*. Cambridge University Press.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Furnas G. W., Landauer T. K., Gomez L. M., and Dumais S. T. (1987). The vocabulary problem in human-system communication: An analysis and a solution. *Bell Communications Research*.

Heidorn, G. (1999). Intelligent writing assistance. In Dale, R., Moisl, H., and Somers, H. eds., *A Handbook of Natural Language Processing Techniques*. Marcel Dekker.

Koppel, M., Argamon, S. and Shimoni, A. (2002), Automatically categorizing written texts by author gender, *Literary and Linguistic Computing* 17(4): 401-412.

Mosteller, F., and Wallace, D.L. (1964). *Inference and disputed authorship: The Federalist. Reading*, Mass.: Addison-Wesley.