How Incorrect Information Delivers Correct Search Results: A Pragmatic Analysis of Queries

Jin Ha Lee (jinlee1@uiuc.edu), Allen Renear (renear@uiuc.edu)
University of Illinois at Urbana-Champaign

Introduction

The accuracy of the information that users bring to the search has been regarded as of primary importance for successful search results in studies of catalog uses and reference service [2]. Especially for known-item searches in which users are trying to find a particular object [7], users' incorrect information about the sought object is considered a serious problem, as even a single non-matching term combined using "and" in a Boolean search statement will result in a search failure [1]. Incorrect information in users' queries is therefore typically regarded as flaws or errors that need to be corrected in order to make the search successful. It is astonishing then, to observe that although (known-item seeking) music queries are frequently riddled with error they are nevertheless more often successful than not. How can this be so if incorrect information is the obstacle it is thought to be on standard accounts of information seeking?

In this paper, we suggest an alternative perspective based on our analysis of real-life queries. We argue that certain kinds of incorrect information can in fact be useful for particular search tasks, and propose that we should advance our understanding of how this works and how a better understanding of this phenomenon might improve search. We draw on real-life examples of human-intermediated music queries to illustrate our case, emphasizing the importance of the role of *pragmatics* of queries. (By pragmatics we refer to the area in contemporary linguistics that includes such topics as presupposition, perfection, conversational implicature, speech acts, and referential use of descriptions.) We show how findings from these areas help explain why it is that queries which appear to be semantically flawed are nonetheless successful, providing at least part of the explanation for why social networking and human mediated approaches often succeed where traditional approaches fail.

A Query Example

On music related forums and Q&A websites, music queries (expressed in natural language) are posted, and then other knowledgeable forum participants proceed to answer them. Without the constraints of having to formulate formal search statements, a wide variety of information is typically provided by the inquirer. Queries presented to intermediaries often contain information that is ambiguous, vague, underspecified, incomplete, false, and sometimes even inconsistent – and yet they perform surprisingly well in eliciting correct answers in human intermediated searches. There are of course a number of well-known retrieval strategies that are designed to accommodate such problems — approximation and fuzzy searching, error correction, disambiguation, authority control, controlled vocabulary, spellchecking, and so on. All of these techniques appear to be based more or less on the view that such queries are simply semantically flawed by inaccurate, approximate, and ambiguous information, requiring various kinds of "correction". Our analysis of these queries and responses suggests a different view, showing that human intermediaries responding to the queries are not simply employing analogues to correcting and approximating retrieval strategies, as demonstrated in a real-life example[i] below:

> [Q.] I need to find the title and artist of a song I heard on the radio. I only caught some of the lyrics. I only heard it once. I am not sure that the wording is absolutely correct. It was a female singer, very plain, beautiful, slow. "there will be no black flag on my door" "I am in love" "I will go down with vengence [sic.]"

The correct answer[ii] to this query is shown below. Although the information provided by the user was partially incorrect, the intermediary was still able to find the right song by matching the incorrect lyrics information appearing on one of the websites for sharing misheard song lyrics.

> [A.][…] I'm pretty certain that the song you heard was Dido's "White Flag". A snippet of the lyrics are:
> "I will go down with this ship/And I won't put my hands up and surrender/
> There will be no white flag above my door/I'm in love and always will be"
> […] Others have also mistaken sections of the lyrics of certain Dido songs:
> Misheard Lyrics Arranged by Artist -> Dido URL: http://www.amiright.com/misheard/artist/dido.shtml
> […] Search Strategy (on Google): "be no black flag" / dido "white flag" lyrics / dido […]

Pragmatic Analysis: Two uses of definite descriptions

To provide an explanation for such success, we apply a well-known distinction in pragmatics, the attributive/referential distinction, initially presented by Keith Donnellan [4]. Although studies have applied Donnellan's distinction to computational models of reference and discourse [3][5][9], there has been, with the exception of Ng's adaptation of Habermas [8], little application of pragmatics to actual information retrieval (IR) interactions.

Donnellan argues that there are two uses of definite descriptions such as "the man drinking a martini". When used *attributively* the referent of the description is whoever or whatever fits that description. However, when used *referentially* the definite description is merely a device for calling attention to some person or thing — which may or may not fit the description in question. Here is Donnellan's familiar example:

> Suppose one is at a party and seeing an interesting-looking person holding a martini glass, one asks, "Who is the man drinking a martini?" If it should turn out that there is only water in the glass, one has nevertheless asked a question about a particular person, a question that it is possible for someone to answer…

There will still be a temptation to see what is going on here as simply a failure of accuracy or precision that is "corrected" by the human intermediary. This temptation, however, must be resisted. For one thing, there may be no correction at all: the human intermediary may have no reason to believe that the description fails to uniquely identify even when it does fail, or they may believe that it fails to uniquely identify, but have no grounds for improvement. In addition, correction may actually reduce the likelihood of a successful outcome, as it is sometimes easier to successfully identify the referent through the incorrect description. But most significantly, as Donnellan points out, the referential use of definite description can be seen to be a distinctly logical feature of some uses and not others and is thus entirely independent of descriptive accuracy *per se*. Donnellan offers this example of a use that is ineluctably attributive:

> [...] we are told that someone has laid a book on our prize antique table, where nothing should be put. The order, "Bring me the book on the table" cannot now be obeyed unless there is a book that has been placed on the table. There is no possibility of bringing back a book which was never on the table and having it be the one that as meant [...]

Compare: "What was the first song the Beatles ever recorded?" This also cannot typically be understood referentially. However in the case of "I am looking for the first song the Beatles recorded..." we cannot determine whether attributive or referential use is intended until we have more context. Contrast these two completions of that query: "...because we have a bet on it." and "...it is something about a Hardee's night." The query with the first completion requires an attributive reading, and can only be satisfied by "Love Me Do" whereas the query with the second completion allows a referential reading and can, at least in some contexts, be satisfied by "It was a Hard Day's Night".

Pragmatics and the Importance of Context Metadata

We can now explain some of the limitations of IR systems with this observation: *current IR systems treat all definite descriptions as attributive, even when they are intended as referential*. And as Donnellan notes, "when a definite description is used attributively in a command or question and nothing fits the description, the command cannot be obeyed and the question cannot be answered." On the other hand a human intermediary easily identifies and accommodates referential use, as in the example[iii] below:

> [Q.] Looking for a song. They lyrics go "My Mamma done told me, When I was in knee-socks" It's a jazzy number.
>
> [A.][…] Well, you were close. The lyrics refer to "knee pants" not "knee socks" and the genre is blues rather than jazz. In fact, the tune is called "BLUES IN THE NIGHT". And yes, it's a very cool song to be sure […]

Here the intermediary understood at least part of the inquirer's description as referential (i.e., "knee-socks" and "jazzy") and was able to find the correct answer[iv]. Understanding that people often confuse Jazz with Blues (and vice versa) may have helped the search in this particular case. However, IR systems that treat all descriptions as attributive will return no results if genre information was used in conjunction with the lyric information in the search statement, since there is no "Jazz" song that satisfies other condition. Unlike retrieval software, humans routinely exploit a vast background of social facts and conventions which we have elsewhere referred to as *context metadata* [6]. Included in this context metadata are the conventions of language, not only the attributive/ referential distinction, as described above, but other discourse conventions, and common knowledge and expectations as well (e.g., performance histories considered to be common knowledge). Incorrect information such as the commonly made mistakes is particularly an interesting kind of *context metadata*. In the context of music IR, consistently misheard lyrics, wrong genre or artist, inexact released dates are more common than not. Some questions that need to be further explored are:

- Which features can be used to make system distinguish between the two types of descriptions in particular searches?
- How can we systematically collect, store, and organize the referential descriptions including common misinformation?
- How do we incorporate the incorrect information about the objects without compromising the information quality?
- How can we safely support inferencing with referential description?

Conclusion

Unlike traditional tools for information access such as indexes, catalogs, or reference materials, the Web contains a significant amount of information that is either partially or completely inaccurate, or has nothing to do with the accuracy (e.g. false beliefs, personal thoughts, opinions, communication, and memories). Finding and exploiting the patterns hidden in the seemingly chaotic mass of incorrect information will be the key for advancing the state-of-the-art in IR. Understanding how queries which appear to be semantically flawed but referentially effective are navigated by human intermediaries provides insight into how we successfully interact with each other around information objects — even though our communication appears to be a tissue of falsehood, ambiguity, and confusion. These insights not only suggest new strategies for improved searching, but, more importantly, by emphasizing the profound and subtle significance of cultural context in the interpretation of queries, they open entirely new opportunities for understanding how our navigation of information objects is determined by shared expectations and conventions.

References

[1] Allen, B. (1989). Recall cues in known-item retrieval. *Journal of the American Society for Information Science, 40*(4), 246-252.

[2] Baker, S. L., & Lancaster, F.W. (1991). *The measurement and evaluation of library services* (2nd ed.). Arlington, VA: Information Resources Press.

[3] Birner, B. J. (1991). Discourse entities and the referential/attributive distinction. Paper presented at the 65th Meeting of the Linguistic Society of America, Chicago.

[4] Donnellan, K. S. (1966). Reference and definite descriptions. *Philosophical Review*, 75, 281-304.

[5] Kronfeld, A. (1986). Donnellan's distinction and a computational model of reference. In *Proceedings of the 24th Meeting of the Association for Computational Linguistics*, 185-191.

[6] Lee, J. H., Hu, X., & Downie, J. S. (2005). Q&A websites: Rich research resources for contextualizing information retrieval behaviors. In *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context* (pp. 33-36).

[7] Lee, J. H., Renear, A., and Smith, L. C. (2006). Known-item searching: Variations on a concept. In *Proceedings of the 69th ASIS&T Annual Meeting*, Austin, Texas.

[8] Ng, K. B. (2002). The applicability of universal pragmatics in information retrieval interaction: A pilot study. *Information Processing and Management*, 38, 237-248.

[9] Onishi, K. H., & Murphy, G. L. (2002). Discourse model representation of referential and attributive descriptions. *Language and Cognitive Processes*, 17, 97-123.

---

[i] From Google Answers (http://answers.google.com/answers/)

[ii] The inquirer verified that the answer is correct in his/her rating feedback.

[iii] From Google Answers (http://answers.google.com/answers/)

[iv] The inquirer verified that the answer is correct in his/her rating feedback.