Helping users understand and recover from interpretation failures in natural language interfaces

Dustin Arthur Smith
MIT Media Lab
E15-358; 77 Massachusetts Ave
Cambridge, MA 02139; USA
dustin@media.mit.edu

Henry Lieberman
MIT Media Lab
E15-320F; 77 Massachusetts Ave
Cambridge, MA 02139; USA
lieber@media.mit.edu

ABSTRACT

When a person determines what he will say and how he will say it, he relies upon beliefs that he predicts to be shared with his audience. The difficulty of this inference task is amplified when a person speaks to a computer, because it is unclear what the computer knows and how the computer will use its knowledge to interpret the speaker's utterance. This lack of transparency and predictably are barriers for natural language interfaces. Despite advances in natural language processing, interpretation failures are not going away. In human-human linguistic communication, ambiguity, vagueness and underspecification are frequent; and the listener is expected to be resilient and flexibly shift between different interpretations when needed, or diagnose when a discrepancy between the speaker and listener's communication goals or beliefs are responsible for the communication failure.

This paper takes the position that augmenting a NLI with non-linguistic modalities may present a mode for recovering from interpretation failures that is richer than relying solely on dialogue. The paper presents a categorization of the levels of information where interpretation failures may occur, and explains the authors' ongoing research using the framework.

Author Keywords

Natural Language Interfaces; End-User Programming

CONTEXT AND MOTIVATION

When people communicate using natural language, we do so in service of specific communication goals. Similar to other purposeful actions, when we speak we try to minimize the effort required to achieve our goals. Fortunately, we can share some of the burden with our audience—in fact, we *expect* our listeners will cooperate. This means they have to share our communication goal(s), and have some way to determine if communication has succeeded. When we communicate with a computer, it's harder, because we don't necessarily know what the computer knows nor how it will interpret our message. This lack of predictability is detrimental to linguistic communication because the content and form of our utterances is influenced by what we expect our audience knows.

When a speaker's message omits information that is essential to its interpretation, the audience is expected to infer what speaker *meant* from what he *said*. From an HCI perspective, the option for users to choose what information to leave out is what gives natural language interfaces their competitive advantage other other methods of supplying information to a computer. The user (speaker) of an natural language interface can choose the information he wants to communicate and the level of detail, leaving the missing information to be filled in by presumably intelligent defaults. The requirement of recovering the intended hidden information poses major challenges for natural language processing: for instance, it is difficult to draw the line between what information was intended to be inferred and what was intended to remain underspecified. A computer interpreter may over-cooperate and make assumptions that we did not expect or think were justified; for instance, imagine if Google Calendar, in addition to assuming that by entering "dinner" you intended the event to occur *tonight*, also chose *who* would join you for dinner!

Wherever there is missing information, there is an inference challenge and room for misinterpretation. Here we summarize the major categories of missing information in natural language:

Divergent Beliefs: The speaker may overestimate the mutual beliefs of the listener, and communicate information that requires background knowledge that the listener does not possess. More specifically, there are many relevant types of knowledge the listener may lack including: the lexical entries (e.g. words), the senses of the words (disjoint semantic effects), and the entities being referenced (**context set**) . Often computers lack the same set of referents: your phone will correctly respond to "call Mom" but not "call her back". Sometimes computers have *too large* a context set: "call John" presents you with 14 guys, some of whom you haven't talked with in years. Note that assumed background knowledge is a function of social identity. If you say "meeting during IAP", you should expect the sense of "IAP" that means a period in the month of January is only known to members of the MIT community. The acronym "MIT" itself—the sense meaning a university in Cambridge, is also a convention shared among a specific set of people; and your decision whether to use this acronym will be influenced by the identifies you are able to impart upon your audience.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

Divergent Communication Goals: Failure can occur at a pragmatic level: namely when the communication goals are incorrectly recognized. Goals can be simple, such as communicating the identity of the entree to a waiter; or very complex: consider the social goal of convincing Susan that Thomas is a dishonest person.

Vagueness: Vagueness is pervasive and can come from words with vague semantics (e.g. "good", "near"), or when a listener hedges the precision of his claim either implicitly (e.g. "I live *100 miles* from here") or explicitly (e.g. "nearly a mile"). With vagueness, the listener is not expected to infer the missing information; as opposed to ambiguity, where the listener is expected to supply information or choose between multiple, conflicting interpretations.

Underspecification: Commonsense assumptions, or what linguists cal **conversational implicatures**, are one of the key sources of missing information in an utterance. They can be thought of as an extreme form of vagueness: when information is missing entirely from the utterance.

Ambiguity: Ambiguity is when the utterance contains missing information that needs to be resolved by the listener, and in doing so, the listener must decide between two or more mutually exclusive interpretations. Structural and lexical ambiguity are two common varieties. Suppose you say "dinner at 6" to Google Calendar. It will create a new event and correctly populate the starting time to 6 pm. However, if you enter "dinner at 9", it will misinterpret the time reference as the inaccurate 9 *am*. If you say this to a person, he will resolve the lexical ambiguity by cooperatively assume that you mean 9 *pm* and not *am*.

To further complicate the matter, the listener is expected to interpret and re-interpret the speaker's utterance *on-line*, revising old assumptions if they are inconsistent with new information [1]. Continuing the previous "dinner at 9" example, our listener would revise his assumptions if you gave him additional, conflicting information, for example that "9" is the name of a nearby restaurant.

We hope the above examples have convinced you that interpretation failures, particularly at the pragmatic level, are an inescapable part of the communication process. Currently, when a computer interprets natural language, the human user is excluded from the interpretation process, or can only interact using a single modality (natural language: spoken or written). How can users interact with these underlying assumptions? How can we make it easier to repair interpretation failures when they inevitability occur, ideally so that the computer does not make the same mistake again? How should we reveal the assumptions involved with language interpretation to end-users in a way that is analogous to a good debugging tool, which allows programmers to quickly track down and correct the root cause of a problem?

THE ANATOMY OF AN INTERPRETATION FAILURE

A person may need to access the following pieces of information to accurately predict the interpretive abilities of a NLI:

1. **What the interface knows**, including:
   (a) the words it knows
   (b) the senses of the words it knows
   (c) the entities it can possibly refer to (*the context set*)
2. **How the interface derived the interpretation**, including:
   (a) the senses of the user's words it chose
   (b) the syntactic structures it inferred
   (c) the pragmatic assumptions it assumed
   (d) the subset of semantic structures (from 1c) it referenced (*the target set*)

All of these pieces of information are only important when the speaker and listener have collectively failed to achieve the shared communication goal; namely, in all cases but when the speaker believes his audience has demonstrated sufficient understanding of the intended message, meeting some **grounding criteria** [2] that indicates mutual belief has been established. Showing an end-user all of this information at once would certainly be distracting.

Meeting the grounding criteria means the interpretation was successful. When the grounding criteria is not met—which is common, it is an opportunity to repair knowledge to improve the interpreter, provided that the user is able to work with the background data in (1a-c) and data derived from the procedures of (2a-d). In a later section, we describe our current research on an event communication task where there is a clear grounding criteria.

The taxonomy also allows us to more explicitly differentiate *types* of interpretation failures, which each may require their own unique mode of interaction. For example, in the context of interpreting referring expressions [5] one way to address the problem of opaque referents (1c) is by displaying what target set has been derived from the context set during the interpretation process (steps 2a-d).

Our research has focused on recovering from pragmatic level interpretation failures (specifically 1c; 2c-d); and, in order to do so, we are building an interface that gives novice users control over these parts of the interpretation process. For example, to address the problem of determining which referents are *plausible* from all the referents that are *possible*, we need to negotiate the assumptions that constraint the context set (1c) with the user (2c). We have built an interface that shows users a visual representation of what assumptions have been made, and allows them to add new assumptions and edit the conditions under which these assumptions are triggered.

The main virtue of an NLI is that a speaker doesn't have to include all details in the message; she can rely on the interpreter to unpack all of the information she intended to convey. This transfers the burden of using assumptions to resolve ambiguities and fill in missing details onto the interpreter. Until NLP is fully capable of dealing with this burden, we take the position of [3] that the problem is best solved through a multimodal approach which combines the contributions of a natural language interface with traditional, direct-manipulation interface. Said another way, multi-modal interactions can provide "training wheels" for developing natural language interfaces.

RESEARCH QUESTIONS

Our research goal is to uncover the best strategies for recognizing and repairing language interpretation failures; which has led us to these questions:

1. When should we choose to leave an interpretation vague, and when should the interpreter intervene by making assumptions or issuing a clarification request? How can we involve the user in this decision?

2. How does the interpreter's transparency affect the interface's overall usability?

3. Does having control over the interpretation decisions help the user achieve the communication goal more quickly?

ONGOING RESEARCH

We have created an interface with the goal of responding to and recovering from semantic and pragmatic communication failures. There are two distinct contributions: The first is an implemented system that both interprets and generates English referring expressions, which came out of the view that language generation can be modeled as a belief-state planning problem, and interpretation as the problem of goal recognition. Our second research effort is to evaluate natural language interfaces that have been extended with direct manipulation interfaces in their effectiveness for allowing end-users to detect and recover from different types of communication failures.

Using a model of language generation and interpretation based on planning and plan recognition, we capture, through user contributions, word definitions and commonsense assumptions– and we represent both as belief-changing actions. Using visualizations and a direct manipulation interface, users can access the interpretation status, inspect which assumptions were made, and suggest or modify existing assumptions. With the aim of providing the functional equivalence of the negotiation stage in interpersonal dialogue, we evaluate the interface by how it allows users to revise and extend assumptions toward successful interpretation.

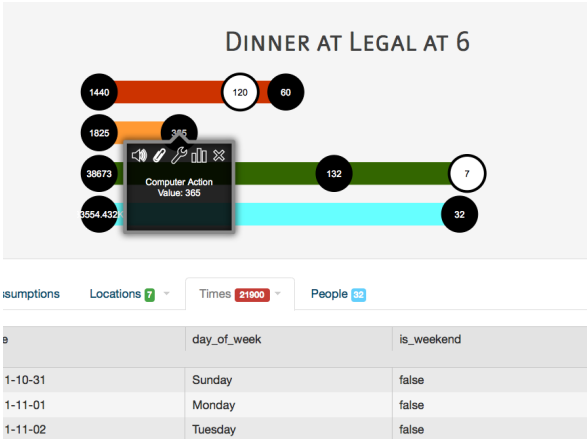

Figure 1. A prototype of the interface, showing the overall progress toward the interpretation goal.

The interface presents the user with a visualization of the interpretation status with respect to the communication goal. When there are multiple valid interpretations, the user can indicate which one he prefers. When a user clicks on an action (represented as a circle) they are given the option to remove or revise it. After modifying an action, the interface will adjust the contrast set and change the status of the overall interpretation process, removing or adding words to the input description if necessary. When the user edits the contrast set or target set, the system determines if this action conflicts with previous assumptions, and otherwise adds it to the interpretation and adjusts the corresponding interpretation status.

Task: communicating English event descriptions

Building on the ambition of our earlier commonsense calendaring application [6], we have situated our language interpretation task in the calendaring domain, where the communication goal is to convey a concrete event representation. The job of the interpreter is to retrieve the same concrete event that the user intended to convey with her description; moving it into the common ground. We represent the result of an event interpretation using four **event components**; and so our communication task can be translated into four selection tasks:

1. Picking a **starting time** from a set of 1440 minutes in a day.

2. Picking a **starting day** from the set of $365n$ days in the next $n = 5$ years.

3. Picking a **location** from the set of all locations in a Point of Interest database, e.g., Yahoo! Local API.

4. Picking a **group of people** from the set of 2^n subgroups of an n-person social network, e.g. Facebook or Google Contacts.

Any combination of multiple choices (subsets) for each component forms a potential **belief state**, which initially represents all of the possible events that can be communicated. For our task domain, the collaborative goal is for the computer and user to establish a definite reference for an event's starting time, starting day, location and attendees. Communication progresses from a state of maximal uncertainty about each of the event's components, to complete certainty, when each event component contains one member, and thus there is only one event in the belief state (overall size = $1 = 1 \times 1 \times 1 \times 1$).

Users can communicate the event using two modalities: written natural language and direct manipulation. Analogous to a direct manipulation selection task, the linguistic task of individuating specific items from a set is called **generating** a **referring expression** [5]. To do this, the user must first select words that discriminate the intended referents from the possible referents. For example, the referring expression "tomorrow" selects an individual starting day from a set of possible days. The intended referent, e.g. November 8th, is called the **target set** and is selected from possible referents called the **context set**, whose non-target set members comprise the **contrast set**.

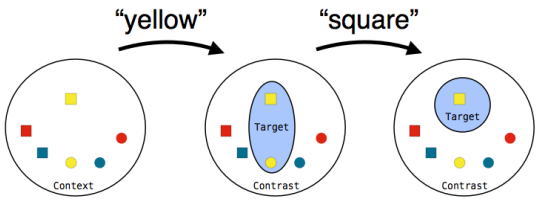

Figure 2. Starting with a *context set* (leftmost circle) containing 6 members, the referring expression "yellow square" incrementally generates a *target set* with one member and a *contrast set* with 5. The progression from left to right shows two actions segments of the interpretation process.

For our problem domain, referring expressions are concerned with localizing specific times, locations and participants for events. Examples of **starting time referring expressions** include: "at noon", "at 6pm" and "after dinner". Because our event communication task is constrained, it gives us clear criteria of when the communication goal has been met:

Grounding Criteria: Does the set of referents (*target set*) specify a concrete event description?

In this view, a single interpretation of an event description is a sequence of interleaved belief state refinement actions, which can either be caused by the user, using language or direct manipulation, or by the interpreter putting forth assumptions. This is conveyed in figure 3:

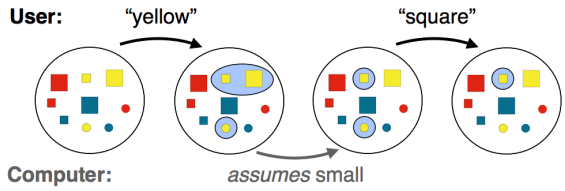

Figure 3. Here we have added three additional objects (large squares) to the context set of figure 2, and show an example of the interpreter cooperatively adding an assumption toward the joint goal deriving an interpretation.

In figure 3, the interpreter offers the assumption that the user intended to specify *small* items. For concreteness, imagine that these are sweets on a dessert tray, and a friend is asking you to pass him one. You know that your friend is on a diet, so when he begins to describe a sugary dessert item, saying "yellow. . . ", you cooperatively provide the assumption that he wants a smaller portion.

Each action, either corresponding to a word or an implicit assumption, is a step in the belief state plan toward achieving the communication goal. Following [1, 4], controlling which assumptions and word-producing actions are deployed can be construed as a planning problem, where the goal is to find the interpretation path(s) with the lowest cost.

Any interpretation is a specific plan, and we use a subway map as a visual metaphor to represent the plan's progress (figure 4). As the plan progresses toward the grounding criteria, we extend the line from starting belief state (containing the contrast set) to the right, a goal state, where the set size is 1.

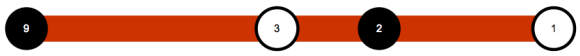

Figure 4. This subway visualization shows the interpretation plan from figure 3 that a user could to see the status of the interpretation task and alter its decisions. Circles represent action's effects and the current size of the target set after the action. User actions, denoted by white circles ③, are observed words or interface actions; assumptions that the computer made are filled circles ❷. The numbers indicate the size of the current target set, progressing from left to the rightmost goal state: ❶ or ①.

The next step in this ongoing research is to conduct a large-scale user experiment to answer the questions described in the previous section. The long term goal of this research is to lower the cost for interacting with an NLI so end users can contribute to the linguistic interpretation process, we can collect culture-specific lexical and semantic knowledge directly from the members of the cultural group who possess it. This knowledge is essential for the pragmatic task of deriving what a speaker *meant* from what they *said*.

BIOGRAPHY

Dustin Arthur Smith, MSc is a Ph.D. student in the Media Lab, working at the intersection of planning and natural language processing. His advisors are Henry Lieberman and Marvin Minsky. Dustin's research is on building easy-to-use language debugging interfaces with the ultimate goal of overcoming the lexical and commonsense knowledge acquisition bottleneck.

Dustin earned his BSc in Computer Science (with a minor in Neuroscience) at Wake Forest University in 2005. He earned his MSc in Media Arts and Sciences from MIT in 2007 with the thesis *EventMinder: A Personal Calendar Assistant That Understands Events*.

REFERENCES

1. Benotti, L. *Implicature as an Interactive Process*. PhD thesis, Université Henri Poincaré, INRIA Nancy Grand Est, Francia, 2010. Examined by Patrick Blackburn, Nick Asher, Bart Geurts, Alexander Koller and Claude Godart.
2. Cahn, J. A psychological model of grounding and repair in dialog. *AAAI Symposium on Psychological Models of Communication in Collaborative Systems* (1999).
3. Cohen, P. The role of natural language in a multimodal interface. In *Proceedings of the 5th annual ACM symposium on User interface software and technology*, ACM (1992), 143–149.
4. Hobbs, J., Stickel, M., Appelt, D., and Martin, P. Interpretation as abduction. *Artificial Intelligence* (1993).
5. Reiter, E., and Dale, R. *Building natural language generation systems*. Cambridge University Press, New York, NY, USA, 2000.
6. Smith, D., and Lieberman, H. *Recognizing and using goals in event management*. PhD thesis, 2009.