On Trying to Do Things with Words

Another plan-based approach to speech act interpretation

Michael J. Hußmann
Heinz Genzmann[1]

FB Informatik, Universität Hamburg
Rothenbaumchaussee 67–69
D-2000 Hamburg 13, West Germany
e-mail: hussmann@rz.informatik.uni-hamburg.dbp.de

Abstract

Usual plan-based approaches to speech act interpretation require that performing a speech act implies its success. These approaches are thus useless for describing failing illocutionary or perlocutionary acts. We propose an alternative plan-based view of speech acts centered around the notion of *trying to do* – as opposed to actually doing – an action. This approach is contrasted with that of Perrault which aims to overcome similar problems.

1. Introduction

The plan-based approach to the analysis of natural language dialogues, inspired by the work of Austin [Austin 62] and Searle [Searle 75] [Searle/Vanderveken 85], has dominated most of the dialogue-oriented NLU research since the late seventies, cf. [Allen 79], [Allen 84], [Cohen 78], [Cohen/Perrault 79], [Perrault/Allen 80], [Litman 85], [Pollack 86]. It is characterized by the following assumptions:

- Utterances are actions planned and executed by the speaker to satisfy some of her goals.
- Speech acts can be represented as operators in planning systems such as STRIPS [Fikes/Nilsson 71] which relate actions to their preconditions and effects.
- The speaker's beliefs and intentions can be inferred by observing her utterances.

Several variants of describing actions in planning systems have been proposed in the literature. We will use the following conventions: an action is defined by a quintuple comprising

- an **action description**, e.g. (pick-up ?x) in the blocks world domain,
- a set of **preconditions**, e.g. (on-table ?x),
- the **add-list**, a set of propositions which become true once the action has been performed, e.g. (holding ?x),
- the **delete-list**, a set of propositions which are no longer true after the action has been performed, e.g. (on-table ?x),
- the **body**, a list of lower-level actions comprising the described action, e.g. the movements of a robot arm in the blocks world domain.

The relationship between these elements is interpreted as *conditional generation* (following [Goldman 70], [Pollack 86]): the execution of the body actions generates the defined action (and thus its effects as described in the add- and delete-list) *iff* the preconditions hold.

All three acts involved in making an utterance, the locutionary, illocutionary, and perlocutionary acts, may fail: the addressee may not hear the utterance, she may not understand the speaker, or she may not react according to the speaker's intentions.

But when actions are defined in terms of their effects, a failure to achieve the effects implies a corresponding failure to perform the action. When I try to drive a nail into the wall with a hammer, and fail, then I *have not driven a nail into the wall*. Similarly, when I utter a declarative sentence, and fail to convince the hearer, I have failed to perform the intended perlocutionary act. When the addressee does not understand me, I have not even performed an illocutionary act, and so on. Regardless of the level of description there's always the chance that the action may fail, i.e. that there was *no* action. Even granted that we could capture practically all of the relevant cases by describing the action on the level of, say, producing sounds, there's no way to relate that level of description to the intentions of the speaker.

This is a rather unfortunate result, as we take the observed action as the starting point for inferring the speakers beliefs and intentions — which may well be the same, regardless of the speech act's success. Clearly, an approach facilitating a uniform treatment of succeeding and failing speech acts would be most welcome.

There are two ways to cope with failing speech acts. The first amounts to weakening the inference from the performance of an action to its effects being achieved, i.e. making it defeasible. The other solution is to describe actions in terms not presupposing their successful execution. The former approach was proposed by Perrault [Perrault 87] and further refined by Appelt and Konolige [Appelt/Konolige 88], while the latter is the one we use.

2. Perrault's approach

Perrault describes the way assertions influence peoples beliefs by distinguishing between *axioms* describing strong evidence for beliefs, and *default rules* capturing the effects of (sometimes failing) speech acts. The most important axioms are:

[1]This paper was written by the first author but describes work jointly undertaken with the second.

Memory: $\vdash B_{x,t}p \supset B_{x,t+1}B_{x,t}p$

Persistence: $\vdash B_{x,t+1}B_{x,t}p \supset B_{x,t+1}p$

Observability: $\vdash DO_{x,t}a \,\&\, DO_{y,t}Obs(x) \supset B_{y,t+1}DO_{x,t}a$

Agents remember their previous beliefs (*memory*), they stick to what they believe they believed previously (*persistence*). If an agent observes another agent, she knows of all actions the observed agent performs (*observability*). *Memory* and *persistence* together imply that agents never forget and never change their beliefs. Observing is regarded as the only dependable mode of acquiring new knowledge.

Speech acts are considered to be a weaker kind of evidence, and thus the effect of uttering a declarative sentence is modelled by default rules:

Declarative rule: $DO_{x,t}(p.) \Rightarrow B_{x,t}p$

Belief transfer: $B_{x,t}B_{y,t}p \Rightarrow B_{x,t}p$

i.e., an agent believes what she thinks other agents believe, provided this is consistent with her previous beliefs (*belief transfer rule*). Uttering a declarative sentence implies by default that the speaker believes its propositional content (*declarative rule*).

For the most part, this theory makes correct predictions. For example:

- The hearer will not be convinced by an assertion if it contradicts one of her previous beliefs.
- A liar will not be convinced by her own lie, but may still believe that she successfully deceived the unsuspecting hearer.

In both of these cases do the axioms (*memory* and *persistence*) override the defaults.

However, adopting Perrault's solution has the unfortunate side effect of depriving speech act rules of their definitional import: an action may be executed without its effects being achieved. What are the effects anyway? According to Perrault, uttering a declarative sentence implies (by default) that the speaker believes its propositional content — this can hardly be thought of as an *effect* of the speech act, and it isn't a precondition, either. As neither the effects of an assertion nor its constituting body actions are specified, this leaves assertions as a primitive, undefined notion. But what is a theory of assertions worth if it does not say what an assertion *is* or what would count as "making an assertion"?

There's also another, more technical problem: when a speaker has no belief whatsoever about P, she can convince herself that P is true by simply uttering "P.": as it's perfectly consistent for her to believe P, both of the declarative rule and the belief transfer rule are applicable and will lead to her believing P. Thus, a speaker may convince herself of anything she is incompetent of.

As Appelt and Konolige have shown, this deficiency can be overcome by employing a more sophisticated nonmonotonic theory, cf. [Appelt/Konolige 88]. But is this added complexity really necessary? Even though Appelt and Konolige claim (without proof) that there can be no specification of the effects of an assertion that applies under all possible circumstances, we aim to achieve just that. Instead of assuming that speech acts sometimes fail to achieve their normal effects, we admit that in fact no (successful) speech act was performed in these cases and relate the speakers behaviour to the intended act by some other means, namely by making explicit the notion of "trying".

3. Trying to do things with words

The missing link between an agent's intentions and her (sometimes unsuccessful) performance of the intended action is the notion of "trying": when an agent has a *present directed intention to A* (cf. [Bratman 87]), she will *try to do A*, and — if the preconditions for *doing A* are satisfied — she will thereby *do A*. Fig. 1 illustrates this:

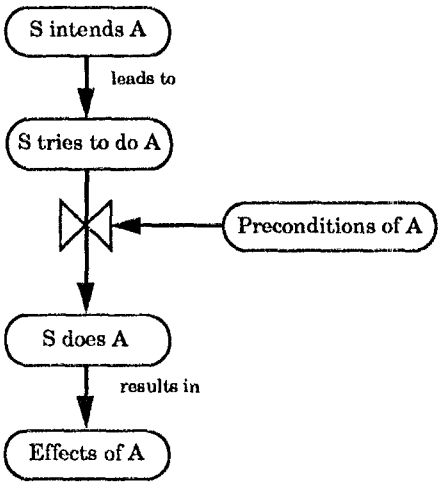

Fig. 2: From intention to action

The *try-to-do* level of description provides an ideal basis for analyzing the agent's intentions, as such an analysis is independent of the action's success.

The interpretation process uses a default rule to determine whether an action was executed successfully:

> **From *try-to-do* to *do*:** If it is known that an agent *S tries-to-do A* in the situation SIT_A[2], then assume (by default) that *S does A* in SIT_A.

The consequence of this conclusion is modelled by another (*non* default) rule:

> **From action to effect:** If an agent *S does A* in the situation SIT_A, then the preconditions of *A* are satisfied in a situation SIT_P temporally including SIT_A (because otherwise it would have been impossible to *do A*), and the effects of *A* are satisfied in a situation SIT_E temporally met by SIT_A.

If this conclusion contradicts a previous belief, the default assumption that *S* had successfully *done A* is defeated.

Here lies one of the main differences between Perrault's approach and our's: where he uses defaults for the inference from the performance of an action to its effects, we use defaults for the inference from *try-to-do* to *do*. The results are similar, but this move allows us to

[2]Perrault associates beliefs and other propositional attitudes with time points, whereas we use situations as partial descriptions of the world over a time interval. For the purpose of this paper we deliberatly slur over this distinction as it is unimportant for the issue at hand.

use strict definitions of speech acts. For example, the speech act *assertion* is defined as follows:

Action: Assert P (in SIT_U)

Preconditions: The hearer believes that the speaker is sincere and competent in SIT_U.

Add-list: The hearer believes that the speaker believes P.

Body: Utter "P." (in SIT_U)

The effect (as specified in the add-list) is what we take to be the illocutionary point: to achieve that the hearer believes that the speaker believes the propositional content.

Let us assume the following scenario: a speaker S utters "P." (referring to a situation SIT_R) in the situation SIT_U. Sincerity and competence of the speaker can be judged by an observer O (who may be identical to the speaker) using the following rules:

Insincerity of a speaker in uttering an assertion: If it is known to O that for some situation SIT_1 temporally included in SIT_R, S believes P to be false in SIT_1, then O knows that S is insincere in SIT_U.

Incompetence of a speaker in uttering an assertion: If it is known to O that for some situation SIT_2 temporally including SIT_R, S has no belief regarding the truth of P in SIT_2, then O knows that S is incompetent in SIT_U.

To complete the picture, we use a belief transfer rule quite similar to the one Perrault uses, except that it blocks inferences of the form

$$B_x B_y B_x p \rightarrow B_x B_x p \rightarrow B_x p$$

which we think are unreasonable: an agent does not adopt the beliefs she thinks other people have of her — at least not automatically. Fig. 2 illustrates the overall structure of the interpretation process, which is the same for the speaker, the hearer, or any incidental overhearer:

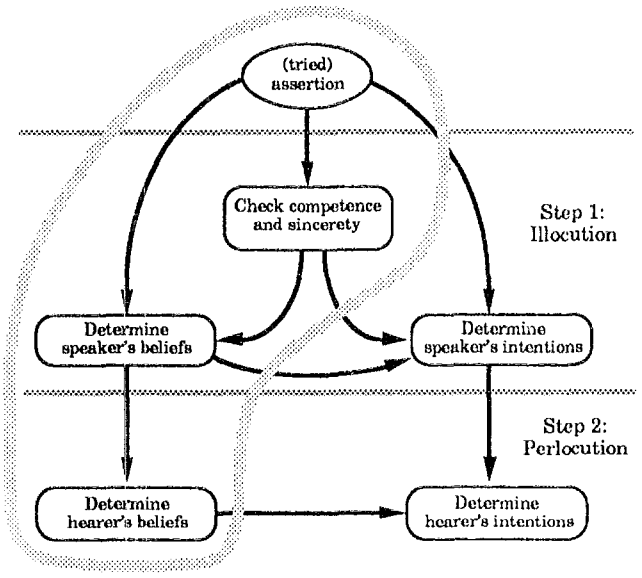

Fig. 2: Interpreting an assertion

The thick gray line encloses the aspects of the interpretation process described in this paper.

An example: the speaker S and the hearer H originally both believe that $\neg P$. There are no other relevant beliefs. When S utters "P.",

- S recognizes herself to be insincere and thus her assertion — as an utterance to herself — fails. On the other hand, it is consistent for S to believe that H will take her to be competent and sincere by default, and that as an utterance to H the assertion will succeed, i.e. $B_S B_H B_S p$ and thus $B_S B_H p$.
- H believes S to be competent and sincere by default and therefore the illocutionary act succeeds as an utterance to H. H believes that S believes P ($B_H B_S p$), but sticks to her previous belief that $\neg P$.

As a result, both S and H will continue to believe $\neg P$, but will also (wrongly) attribute to the other a belief that P. If instead the speaker had no belief about P initially, she would judge herself incompetent and again would not trust her own words.

Determining the speaker's intentions is considerably more complex, as it requires search in a web of action descriptions linked by *(conditionally-) generates* and *generated-by* relations. In general, the observed action A_O and the intended action A_I are linked by a path

$$A_O \textit{ generated-by}^* \textit{ generates}^* A_I$$

The separation of the analysis of belief and intentions enables a straight-forward interpretation of a speech act's effects while offering a sound basis for the subsequent analysis of the speaker's and hearer's intentions.

4. Intending to try

We have described *trying-to-do A* as an intermediate step between *intending* to A and *doing* A. An agent may describe her actions as "doing A" or "trying to do A", and the latter will often just reflect her doubts as to whether she will succeed. Nevertheless, if A is an action, then so is *try* A, and *intending* to *try* A is a genuine intention distinct from an intention to A.

In both cases will the intending agent *try-to-do A*, and sometimes *do* A, but even if she fails to *do* A, she cannot fail to *try* A. *Trying* A will sometimes have the effects of *A-ing* — and, if it does, will have caused these effects — but its only necessary result is that A has been tried. Therefore, *trying* A can be successful where *A-ing* is not. For example, I may intend to try to move a heavy log (cf. [Bratman 87:38f]) while strongly believing I will fail — just to demonstrate that I cannot move it. This demonstration is successful only if I do *not* succeed in moving the log despite trying real hard to do so.

Bratman (cf. [Bratman 87:111ff]) did not recognize the difference between *trying-to-do A* as a consequence of *intending* to A and *trying-to-do A* as a consequence of *intending* to *try* A. His term *endeavoring* (adopted from [Chisholm 76]) encompasses both cases. This missing distinction seems to be responsible for most of the less elegant aspects of Bratman's theory, especially the way he relates *acting with an intention* to *acting intentionally*. On the "Simple View", an agents *doing A intentionally* implies her *intention to A*. Bratman dismisses this Simple View on the grounds that an agent will sometimes act intentionally without (in a strong sense) intending the action.

Consider the following example taken from [Bratman 87:137]: Bratman very much wants to marry Susan, and he equally much wants to marry Jane. He knows he cannot marry both but is unable to resolve the conflict. Therefore he hopes that Susan and Jane will settle the issue for him: he tries to persuade both women to marry him, expecting that just one of the two will agree. This kind of behaviour seems perfectly rational — though probably immoral — but according to Bratman's strong consistency requirement for intentions he should not intend to persuade Susan *and* Jane to marry him when he knows he cannot marry both (and thus cannot achieve both perlocutionary acts). Bratman concludes that he intentionally persuades Susan as well as Jane, without endorsing the inconsistent intention to persuade both. Instead he proposes "guiding desires" as a weaker kind of intentions guiding an agents conduct.

We can offer a simpler solution: while sticking to the Simple View, we agree that it is irrational for Bratman to intend to persuade two women to marry him. It is not (necessarily) irrational, however, to intend to *try* to persuade both women, believing that at most one of the two will agree, and such an intention will lead to the same behaviour towards Susan and Jane as an intention to persuade them *simpliciter* would. The distinction between stronger and weaker kinds of intentions is therefore unnecessary.

Returning to the issue of speech act interpretation, consider the following case: Mary is tried for a crime she did not commit. She has all the evidence against her, though, and thus she is convinced no one will believe her if she pleads "Not guilty", or even trust her she believes it herself. Unfortunately, Mary cannot rationally intend to assert she is innocent when she is certain of her failure, so what could she do? Again, the *intention to try* comes to the rescue: poor Mary may rationally intend to try to assert she is innocent (as she has every reason to do so), and this intention will lead to her uttering "Not guilty".

5. What has been achieved?

We have developed a descriptive framework to provide a uniform account of successful as well as unsuccessful speech acts. The notion of *trying-to-do* an action is applicable in both cases and can serve as the basis for analyzing the speaker's intentions. Our theory makes all the correct predictions that Perrault's theory makes, but it does additionally handle the cases which are problematic for Perrault's account.

MEDIAS, a module for speech act interpretation along the lines of the approach presented here has been implemented using the expert system shell HARES as a rapid prototyping tool [Genzmann 89]. MEDIAS handles assertions as well as yes/no questions, distinguishing information-seeking from information-probing questions.

We are currently investigating the role of natural language utterances in initiating, planning, and coordinating cooperative behaviour (cf. [Werner 89]). This research will build upon and extend the first prototype of the MEDIAS system.

References

[Allen 79] James F. Allen: A Plan Based Approach to Speech Act Recognition. Technical Report TR 121/79, University of Toronto, 1979.

[Allen 84] James F. Allen: Recognizing intentions from natural language utterances. In: Michael Brady, Robert C. Berwick (eds.): Computational models of discourse. Cambridge, Massachusetts: The MIT Press 1984. 107—166.

[Appelt/Konolige 88] Douglas Appelt, Kurt Konolige: A Practical Nonmonotonic Theory for Reasoning about Speech Acts. In: Proc. of the 26th Annual Meeting of the ACL, State University of New York at Buffalo, June 1988. 170—178.

[Austin 62] J. A. Austin: How to Do Things with Words. London: Oxford University Press 1962.

[Bratman 87] Michael E. Bratman: Intention, Plans, and Practical Reason. Cambridge, Massachusetts: Harvard University Press 1987.

[Chisholm 76] Roderick Chisholm: Person and Object. LaSalle, Ill.: Open Court 1976.

[Cohen 78] Philip R. Cohen: On Knowing What to Say: Planning Speech Acts. PhD thesis, University of Toronto, 1978.

[Cohen/Perrault 79] Philip R. Cohen, C. Raymond Perrault: Elements of a Plan-based Theory of Speech Acts. In: Cognitive Science, Vol. 3, 1979. 117—212.

[Fikes/Nilsson 71] Richard E. Fikes, Nils J. Nilsson: STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. In: Artificial Intelligence, Vol. 2, 1971. 189—208.

[Genzmann 89] Heinz Genzmann: Eine Untersuchung zur automatischen Modellierung natürlichsprachlicher Dialogstrukturen in der Mensch-Maschine-Kommunikation. Diploma thesis at the Computer Science Department of the University of Hamburg, August 1989.

[Goldman 70] Alvin I. Goldman: A Theory of Human Action. Englewood Cliffs, N.J.: Prentice-Hall 1970.

[Litman 85] Diane Judith Litman: Plan Recognition and Discourse Analysis: An integrated Approach for Understanding Dialogues. Technical Report 170, Dept. of Computer Science, University of Rochester 1985.

[Perrault/Allen 80] C. Raymond Perrault, James F. Allen: A plan-based analysis of indirect speech acts. In: American Journal of Computational Linguistics, No. 6, Vol. 3, 1980. 167—182.

[Perrault 87] C. Raymond Perrault: An Application of Default Logic to Speech Act Theory. Report CSLI-87-90. CSLI, Stanford, California, March 1987.

[Pollack 86] Martha Elizabeth Pollack: Inferring Domain Plans in Question-Answering. Technical Note 403. AI Center, Computer and Information Sciences Division, SRI International, December 1, 1986.

[Searle 75] John R. Searle: A Taxonomy of Illocutionary Acts. In: K. Gunderson (ed.): Language, Mind, and Knowledge. Minneapolis: University of Minnesota Press 1975.

[Searle/Vanderveken 85] John R. Searle, Daniel Vanderveken: Foundations of Illocutionary Logic. Cambridge: Cambridge University Press 1985.

[Werner 89] Eric Werner: Cooperating Agents: A Unified Theory of Communication and Social Structure. In: M. Huhns, L. Gasser (eds.): Distributed Artificial Intelligence, Vol. 2. Morgan Kaufman and Pitman Publishers, 1989.