Abduction, Conversational Implicature, and Misleading (Extended Abstract)

Chiaki Sakama
Wakayama University
sakama@sys.wakayama-u.ac.jp

Katsumi Inoue
National Institute of Informatics
inoue@nii.ac.jp

January 2015

1 Introduction

In conversation or dialogue, people use *abduction* to understand reasons behind utterances. Suppose that your colleague says "I will be at my office this weekend". Then you may surmise that he/she has much work to do. In this dialogue, the utterance is considered an evidence provided by the colleague, then you seek reasons to explain the utterance. You abduce that your colleague would have much work to do if you believe the implication "$much\_work \supset office\_weekend$". Given an utterance by a speaker, a hearer could perform two different types of abduction. The first one is to produce a hearer's belief of a fact that could explain an evidence provided by a speaker. The above example is of this type of abduction. The second one is to produce a hearer's belief of a speaker's belief which could explain the speaker's utterance. In the above example, suppose that you believe that *the colleague believes* "$much\_work \supset office\_weekend$". Then you abduce that *the colleague believes that* he/she has much work to do. In this case, however, you do not necessarily believe yourself that the colleague has much work to do. In this way, a hearer may use abduction not only for generating assumption that accounts for the utterance, but for generating assumption on the belief state of a speaker who makes the utterance.

From a speaker's viewpoint, a speaker often intends a hearer to abduce a fact ψ in face of the speaker's utterance φ but in fact $\neg\psi$ is the case. Suppose that a guest to a party says "I have a good time, thank you" to the host, but it is a flattery and the party is terrible in fact. The guest considers that the host believes the implication "$enjoy \supset good\_time$" and that the host abduces "*enjoy*" in face of the utterance "$good\_time$." The guest believes $\neg enjoy$ but he/she does not want to make the host believe it. As such, a speaker could use abduction for the purpose of *misleading* a hearer. On the other hand, if the host is smart enough to notice that it is a flattery, that is, the guest has said "$good\_time$" but the host believes that the guest does not tell the truth, then the host might consider that "$\neg good\_time$" is the case and *deduce* $\neg enjoy$ by the implication "$enjoy \supset good\_time$."

In conversation or dialogue, the notion of *conversational implicature* [3] is known as pragmatic inference in linguistic phenomena. The principal subject is to investigate

the meaning of a sentence more than what is actually said. For instance, if a speaker utters the sentence "I have two children", it normally implicates "I do *not* have *more than* two children". This is called a *scalar implicature* (or *Q-implicature*) which says that a speaker implicates the negation of a semantically stronger proposition than the one asserted [7]. Generally, a hearer infers $\neg\psi$ from the utterance φ and the implication $\psi \supset \varphi$ by Q-implicature. This is in contrast with abduction in dialogue, however. In abduction, a hearer infers ψ from the utterance φ and the implication $\psi \supset \varphi$. Thus, two inferences appear to reach opposite conclusions in face of an utterance. In conversational implicature, an alternative implicature, called the *I-implicature*, implies a semantically stronger sentence than what is actually uttered. The clash between Q- and I-implicatures has been studied in the field of *pragmatics* that concerns with the meaning of sentences in conversation [1, 5, 7]. Some researchers have pointed out the utility of abduction in interpreting speech acts [4, 6, 8]. To the best of our knowledge, however, relations or differences between abduction and conversational implicature have never been formally explored in the literature.

In this paper we provide a formal account of abduction and conversational implicature in human dialogues. To this end, we use a propositional modal logic of belief and knowledge. We first formulate two different types of abduction in this logic. We next formulate conversational implicature and contrast them with abduction. Misleading by abduction or conversational implicature is also discussed. The results of this paper characterize how hearers use abduction or conversational implicature to figure out what speakers have implicated, and how speakers use them to mislead hearers. The results also show that how abduction is distinguished from conversational implicature.

2 Abduction in Dialogue

We use a propositional modal logic of knowledge and belief that is standard in the literature. A sentence $B_a\varphi$ is read as "an agent a believes a sentence φ", $K_a\varphi$ is read as "a knows φ", and $C\varphi$ is read as "it is common knowledge that φ".[1] It holds that $K_a\varphi \supset B_a\varphi$, $C\varphi \supset \varphi \wedge K_a\varphi \wedge K_aC\varphi$, etc. The logic has the axioms and inference rules of the system KD45 for B_a and S5 for K_a. We assume a dialogue between two agents, called a speaker and a hearer. Each agent has a (consistent) set of sentences as background knowledge and believes those sentences. When a speaker utters a (consistent) sentence, a hearer performs abduction to explain reasons behind the utterance.

Definition 2.1 (objective abduction) Let a be a hearer and b a speaker. When b utters a sentence φ ($\not\equiv \bot$), a sentence ψ is inferred by *objective abduction* (*O-abduction*) from φ by a if

$$B_a\varphi \wedge B_a(\psi \supset \varphi) \wedge \neg B_a\neg\psi. \tag{1}$$

In this case, ψ is called an *O-explanation* of φ. We write $O\text{-}abd_a(\varphi, \psi)$ if ψ is an O-explanation of φ by a.

[1] For the precise definition of $C\varphi$, see [2].

(1) means that a hearer a believes the utterance φ, and a believes the implication $\psi \supset \varphi$ and disbelieves $\neg\psi$ in his/her background knowledge. In this case, a hearer a infers ψ as an explanation for the utterance φ. It is called "objective" abduction because abduction is performed based on the objective fact of an utterance. Note that every explanation is consistent ($\psi \not\equiv \bot$). If $\psi \equiv \bot$, then $\neg B_a \neg\psi$ in (1) is false. There exist multiple explanations ψ for an utterance φ in general, and a hearer selects the best explanation to be believed. We do not address the issue further in this paper.

Definition 2.2 (subjective abduction) Let a be a hearer and b a speaker. When b utters a sentence φ ($\not\equiv \bot$), a sentence $B_b\psi$ is inferred by *subjective abduction* (*S-abduction*) from φ by a if

$$B_a B_b \varphi \wedge B_a B_b(\psi \supset \varphi) \wedge \neg B_a \neg B_b \psi. \tag{2}$$

In this case, $B_b\psi$ is called an *S-explanation* of φ. We write $S\text{-}abd_{ab}(\varphi, \psi)$ if $B_b\psi$ is an S-explanation of φ by a.

(2) means that a hearer a believes that a speaker b believes his/her utterance φ, and a believes that b believes the implication $\psi \supset \varphi$, and a disbelieves that b disbelieves ψ. In this case, a hearer a infers $B_b\psi$ as an explanation for the utterance φ. It is called "subjective" abduction because abduction is performed based on the hearer's subjective view on the speaker's belief state.

In objective abduction, a hearer may believe an O-explanation ψ which accounts for an utterance φ by a speaker. In subjective abduction, on the other hand, a hearer may believe $B_b\psi$ but does not necessarily believe ψ by himself/herself. A connection between O-abduction and S-abduction is as follows.

Proposition 2.1 *$S\text{-}abd_{ab}(\varphi, \psi)$ implies $O\text{-}abd_a(B_b\varphi, B_b\psi)$.*

The differences between $S\text{-}abd_{ab}(\varphi, \psi)$ and $O\text{-}abd_a(B_b\varphi, B_b\psi)$ are twofold. First, a speaker b utters a sentence φ in $S\text{-}abd_{ab}(\varphi, \psi)$, while b utters his/her belief $B_b\varphi$ in $O\text{-}abd_a(B_b\varphi, B_b\psi)$. Second, a hearer a believes that a speaker b believes the implication $\psi \supset \varphi$ in $S\text{-}abd_{ab}(\varphi, \psi)$, while a believes a weaker implication $B_b\psi \supset B_b\varphi$ (if a speaker b believes ψ then b believes φ) in $O\text{-}abd_a(B_b\varphi, B_b\psi)$. One could also use O-abduction to abduce the belief state of a hearer by $O\text{-}abd_a(\varphi, B_b\psi)$ if $B_a\varphi \wedge B_a(B_b\psi \supset \varphi) \wedge \neg B_a \neg B_b\psi$. In this case, a speaker utters a sentence φ and a hearer abduces the belief state of the speaker which explains the utterance. All $S\text{-}abd_{ab}(\varphi, \psi)$, $O\text{-}abd_a(B_b\varphi, B_b\psi)$ and $O\text{-}abd_a(\varphi, B_b\psi)$ abduce the belief state of a speaker in different ways.

Suppose that a hearer a infers an O-explanation ψ or an S-explanation $B_b\psi$ in face of a speaker b's utterance φ. If it is in fact ψ or $B_b\psi$, the hearer successfully understands a reason behind the utterance φ. On the other hand, if it is in fact $\neg\psi$ or $\neg B_b\psi$, the hearer *misunderstands* a reason behind the utterance. In the introductory example, a colleague says "I will be at my office this weekend" and you conjecture "he/she has much work to do" by O-abduction. If you say "You seem to have much work to do" in response to the colleague, and he/she says "Not at all, I just come to my office to surf the net", then you realize your incorrect abduction. As such, abduction is *nonmonotonic* in the sense that an explanation might be withdrawn if it turns out incorrect.

3 Conversational Implicature

Conversational implicature [3] is a pragmatic inference to an implicit meaning of a sentence that is not actually uttered by a speaker. In his *maxims of conversation*, Grice introduces two maxims of quantity:

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

Based on these two maxims, two principles are introduced from the speaker's viewpoint [5, 7]:[2]

Q-principle: Say as much as you can.

I-principle: Say no more than you must.

These two principles correspond to the next implicatures from the hearer's viewpoint.

Q-implicature: Imply the negation of a semantically stronger sentence than what is actually uttered.

I-implicature: Imply a semantically stronger (or more specific) sentence than what is actually uttered.

These two implicatures apparently conflict because "the Q-implicatures induce the *negation* of the very sort of stronger interpretation that the I-implicatures actually appear to be promoting" [7]. Various attempts to resolve the clash are proposed in pragmatics [1, 5, 7]. In what follows, we formulate these two implicatures in our logic.

Definition 3.1 (Q-implicature) Let a be a hearer and b a speaker. When b utters a sentence $\varphi\,(\not\equiv\bot)$, a sentence $B_b\neg\psi$ is inferred by *Q-implicature* from φ by a if

$$B_aB_b\varphi \wedge C\,(\psi \supset \varphi) \wedge \neg B_aB_b\psi. \tag{3}$$

We write $Q\text{-}imp_{ab}(\varphi,\psi)$ if $B_b\neg\psi$ is inferred by Q-implicature from φ by a.

(3) is explained as follows. First, a hearer a believes that a speaker b believes his/her utterance φ. Otherwise, it would be meaningless to infer the implicit meaning behind the utterance. Second, $\psi \supset \varphi$ is not just a private belief of the hearer, but a *common knowledge* that is shared by the speaker and the hearer. Third, the hearer disbelieves that the speaker believes a sentence ψ which is stronger than φ. In this case, the hearer infers that the speaker believes $\neg\psi$. There exist multiple sentences ψ satisfying (3) in general, and a hearer select an appropriate one from them in the context. The I-implicature is defined in a similar way.

Definition 3.2 (I-implicature) Let a be a hearer and b a speaker. When b utters a sentence $\varphi\,(\not\equiv\bot)$, a sentence $B_b\psi$ is inferred by *I-implicature* from φ by a if

$$B_aB_b\varphi \wedge C\,(\psi \supset \varphi) \wedge \neg B_a\neg B_b\psi. \tag{4}$$

We write $I\text{-}imp_{ab}(\varphi,\psi)$ if $B_b\psi$ is inferred by I-implicature from φ by a.

[2] "Q" means *quantity* and "I" means *informativeness*. The I-principle is called the R-principle (*relevance*) in [5].

In both (3) and (4), conversational implicature is based on common knowledge, that is, both a speaker and a hearer know the truth of the implication $\psi \supset \varphi$, and each one also knows that the other party knows the truth of the sentence. The reason of using common knowledge here is explained as follows. In Q-implicature (resp. I-implicature), a hearer believes that a speaker implies $\neg\psi$ (resp. ψ) by an utterance φ. In this case, the hearer knows the implication $\psi \supset \varphi$ and, at the same time, the hearer knows that the speaker knows the same implication. If the hearer does not know whether or not the speaker knows the implication, then the hearer cannot infer the intended meaning of the speaker's utterance. Conversely, a speaker implies $\neg\psi$ (resp. ψ) by an utterance φ in terms of his/her knowledge of $\psi \supset \varphi$. If the speaker does not know whether or not the hearer knows the implication, then the speaker cannot expect the hearer's reasoning by Q-implicature (resp. I-implicature). So if the speaker utters φ, he/she knows that the hearer knows the implication $\psi \supset \varphi$. Thus conversational implicature is in effect if and only if a speaker and a hearer share the same knowledge and each one knows that the other party also shares the same knowledge. [3] The difference between two implicatures comes from a hearer's belief of whether a speaker believes a stronger sentence or not. Given an utterance φ, if a hearer disbelieves that a speaker believes a stronger sentence ψ, then the hearer interprets that the utterance Q-implicates $\neg\psi$. By contrast, if a hearer disbelieves that a speaker disbelieves a stronger sentence ψ, then the hearer interprets that the utterance I-implicates ψ. Since $B_b\psi \wedge B_b\neg\psi \supset \bot$, the conclusions derived by two implicatures contradict with each other. On the other hand, it may happen that $\neg B_a B_b\psi \wedge \neg B_a\neg B_b\psi$, so it is a hearer's option to decide which implicature is to be applied in the context where $B_a B_b\varphi \wedge C(\psi \supset \varphi)$ holds.

Both abduction and conversational implicature infer information behind an utterance. According to our formulation, an essential difference between the two lies in the use of implication $\psi \supset \varphi$. In abduction, it is a hearer's private belief: a hearer believes $\psi \supset \varphi$ in O-abduction while a hearer believes that a speaker believes $\psi \supset \varphi$ in S-abduction. This is because abduction is a process of private reasoning, and one can reason abductively without knowing the belief state of the other party. By contrast, in conversational implicature the implication is common knowledge: a speaker and a hearer share the same implication. This is because conversation aims at communicating information. Since $C\varphi \supset B_a\varphi$ and $C\varphi \supset B_a B_b\varphi$, one may use common knowledge for the purpose of abduction. On the other hand, it is inappropriate to use one's private belief to infer conversational implicature. A hearer's reasoning based on his/her private belief does not always reflect a speaker's intention. Formally, the following relation holds between S-abduction and I-implication.

Proposition 3.1 *$I\text{-}imp_{ab}(\varphi, \psi)$ implies $S\text{-}abd_{ab}(\varphi, \psi)$.*

As mentioned above, one can use common knowledge for the purpose of abduction. This means that a hearer may reach to opposite conclusions by using S-abduction and Q-implicature. Suppose that $B_a B_b\varphi \wedge C(\psi \supset \varphi)$ holds. In this case, if $\neg B_a\neg B_b\psi$ then $B_b\psi$ is inferred by S-abduction. Else if $\neg B_a B_b\psi$ then $B_b\neg\psi$ is inferred by Q-implicature.

[3] One may consider a weaker definition of implicature using "common belief" instead of "common knowledge", but we do not address such alternative definitions here.

4 Misleading

A hearer believes a speaker's utterance ($B_a\varphi$) in O-abduction, and believes that a speaker believes his/her utterance ($B_aB_b\varphi$) in S-abduction. We next consider a dialogue in which a hearer does *not* believe the speaker's utterance.

Example 4.1 Suppose the Turing's imitation game in which a human judge asks questions to an interlocutor in order to determine whether he or she is interacting with a human or a machine. Consider the dialogue in a game.

Judge (a): Are you a machine?

Interlocutor (b): I'm a human.

Suppose that the judge believes the implication: "$\neg\, machine \supset human$". Given the response "$human$" by the interlocutor, will the judge believe that the interlocutor is not a machine (by O-abduction)?

In the Turing imitation game, a machine attempts to convince a judge that it is human through appropriate, and often *deceptive*, responses. In this dialogue, if the judge disbelieves the utterance φ by the interlocutor, then $\neg B_a\varphi$ holds in (1) and the judge does not abduce that the interlocutor is not a machine. What happens if the judge believes the *falsity* of the utterance by the interlocutor? Suppose a dialogue in which a speaker (b) utters φ but a hearer (a) believes the contrary $\neg\varphi$. In this case, it holds that $B_a\neg\varphi \wedge B_a(\psi \supset \varphi) \supset B_a\neg\psi$, and the hearer believes $\neg\psi$ if he/she believes $\psi \supset \varphi$. In Example 4.1, if the judge believes that the interlocutor is not human, "$B_a\neg human$" and "$B_a(\neg\, machine \supset human)$" lead to the conclusion "$B_a machine$". Likewise, it holds that $B_aB_b\neg\varphi \wedge B_aB_b(\psi \supset \varphi) \supset B_aB_b\neg\psi$. So if a hearer a believes that a speaker b believes the implication $\psi \supset \varphi$, and the hearer also believes that the speaker is *lying*, i.e., the hearer believes that the speaker believes the falsity of his/her utterance φ, then a believes that b believes $\neg\psi$. In Example 4.1, if the judge believes that the interlocutor believes the falsity of his/her utterance, and the judge believes that the interlocutor believes $\neg\, machine \supset human$, then the judge believes that the interlocutor believes $machine$. In this way, when a hearer believes the falsity of an utterance φ, the hearer would believe the negation of a sentence that explains φ. Moreover, if a hearer believes that a speaker is lying, then the hearer could infer reasons behind the act (a speaker lies φ to make a hearer believe ψ by abduction, but in fact the speaker believes $\neg\psi$).

From a speaker's viewpoint, a speaker b may believe that a hearer a would abduce ψ as a result of the speaker's utterance φ. In Example 4.1, suppose that the interlocutor is in fact a human. Then consider the dialogue.

Judge (a): Are you a machine?

Interlocutor (b): Shall I sing a song?

The interlocutor expects that his/her response would make the judge abduce the fact that "the interlocutor is a human" based on his/her belief that the judge believes the implication "$human \supset sing$". Thus, a speaker will decide what to say by considering the effect of his/her utterance on the hearer's side. A speaker may use this to *mislead* a hearer to reach a wrong assumption.

Definition 4.1 (misleading by O-abduction) Let a be a hearer and b a speaker. When b utters a sentence $\varphi\,(\not\equiv \bot)$, b *misleads* a *by O-abduction* if

$$B_b\,(B_a\varphi \wedge B_a(\psi \supset \varphi) \wedge \neg B_a\neg\psi) \wedge B_b\neg\psi. \tag{5}$$

We write $O\text{-}mislead_{ba}(\varphi, \psi)$ if b's utterance φ misleads a to abduce an O-explanation ψ.

(5) says that a speaker b believes that his/her utterance would lead a hearer a to an assumption ψ by O-abduction, however, b believes $\neg\psi$. Depending on situations, a speaker may use a slightly weaker version of misleading by replacing $B_b\neg\psi$ with $\neg B_b\psi$ in (5).

In Example 4.1, suppose that the interlocutor is in fact a machine. The interlocutor b believes that the judge a believes the response $\varphi = human$ by b, and b also believes that a believes the implication $\neg\, machine \supset human$ while disbelieves $\neg\psi = machine$. If the interlocutor b has ability to recognize itself and believes that it is a machine $B_b\neg\psi$, the interlocutor misleads the judge by the response φ.

A speaker's utterance will change depending on his/her belief that whether a hearer believes the speaker's utterance or not. In Example 4.1, suppose that the interlocutor is a machine and it considers that the judge will doubt his/her response. In this situation, consider the dialogue.

Judge (a): Are you a machine?

Interlocutor (b): Yes, I'm a machine.

If the judge believes the falsity of the utterance, the judge interprets the contrary of the response and concludes that the interlocutor is a human ($B_a\neg machine \wedge B_a(\neg\, machine \supset human) \supset B_a human$). However, this is what the interlocutor has intended. In this case, the interlocutor reasons by the formula: $B_bB_a\neg\lambda \wedge B_bB_a(\mu \supset \lambda) \supset B_bB_a\neg\mu$. The interlocutor b believes that the judge a believes the contrary of the utterance $\lambda = machine$, expecting that the judge reaches the wrong conclusion $\neg\mu = human$ using the implication $\mu \supset \lambda$.

Misleading using S-abduction is defined in a similar way.

Definition 4.2 (misleading by S-abduction) Let a be a hearer and b a speaker. When b utters a sentence $\varphi\,(\not\equiv \bot)$, b *misleads* a *by S-abduction* if

$$B_b\,(B_aB_b\varphi \wedge B_aB_b(\psi \supset \varphi) \wedge \neg B_a\neg B_b\psi) \wedge B_b\neg B_b\psi. \tag{6}$$

We write $S\text{-}mislead_{ba}(\varphi, \psi)$ if b's utterance φ misleads a to abduce an S-explanation $B_b\psi$.

A weaker version of this is defined by replacing $B_b\neg B_b\psi$ with $\neg B_bB_b\psi$ in (6). By Proposition 2.1, the next result holds.

Proposition 4.1 $S\text{-}mislead_{ba}(\varphi, \psi)$ *implies* $O\text{-}mislead_{ba}(B_b\varphi, B_b\psi)$.

Note that $O\text{-}mislead_{ba}(\varphi, \psi)$ or $S\text{-}mislead_{ba}(\varphi, \psi)$ does not necessarily succeed. A speaker believes that the condition of performing O-abduction or S-abduction is satisfied on a hearer's side. However, this does not mean that the hearer actually performs abduction to explain the utterance φ. Moreover, there is a possibility that the hearer could reach an explanation which is different from ψ or $B_b\psi$. Note also that a speaker does not always utter falsity in misleading. A speaker may utter what he/she believes true while expecting a hearer will make an incorrect abduction. The interlocutor (machine) may respond that "I often make errors" expecting that the judge will consider it a human by the implication "$human \supset error$." The machine in fact often makes calculation errors by programming bugs, however. Such a speech act is often said "indirect lies" or "lying while saying the truth" [9]. Conversational implicature could also be used for misleading hearers in similar ways.

References

[1] Carson, R. Quantity maxims and generalized implicature. *Lingua* 96:213–244 (1995).

[2] Fagin, R., Halpern, J. Y., Moses, Y. and Vardi M. Y. *Reasoning about Knowledge*. MIT Press (1995).

[3] Grice, H. P. Logic and conversation. In: P. Cole & J. Morgan (ed.), *Syntax and Semantics, 3: Speech Acts*, pp. 41–58, Academic Press (1975).

[4] Hobbs, J. R., Stickel, M. E., Appelt, E. E. and Martin, P. Interpretation as abduction. *Artificial Intelligence* 63:69–142 (1993).

[5] Horn, L. R. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: D. Schiffrin (ed.), *Meaning, Form and Use in Context: Linguistic Application*, pp. 11–42, Georgetown University Press (1984).

[6] Janíček, M. Abductive reasoning for continual dialogue understanding. *New Directions in Logic, Language and Computation, Lecture Notes in Computer Science* 7415, pp. 16–31, Springer (2012).

[7] Levinson, S. C. Minimization and conversational inference. In: J. Verchueren, M. Bertuccelli-Papi (eds.), *The Pragmatic Perspective*, pp. 61–129, Benjamins Publishing (1987).

[8] McRoy, S. and Hirst, G. Abductive explanation of dialogue misunderstandings. In: *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pp. 277–286 (1993).

[9] Vincent, J. M. and Castelfranchi, C. On the art of deception: how to lie while saying the truth. In: *Proceedings of the Conference on Pragmatics*, Studies in Language Companion Series 7, pp. 749–777 (1979).