Pacific Association for Computational Linguistics

A COMPUTATIONAL MODEL OF IRONY INTERPRETATION

AKIRA UTSUMI
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, JAPAN
utsumi@utm.dis.titech.ac.jp

This paper presents a computational model for interpreting verbal irony. Its main features are 1) the first implemented model of irony, 2) an underlying comprehensive theory that covers a wider range of irony than previous theories, and 3) consistency with various empirical findings on irony. The algorithm that embodies the model decides whether a given utterance is ironic by measuring to what degree it satisfies linguistic properties of irony and by examining the discourse context for a proper situational setting of irony, and then outputs the speaker's ironic intention.

Key words: irony, nonliteral language, cognitive modeling, pragmatics.

1. INTRODUCTION

Verbal irony is an intelligent, witty usage of language found in ordinary language activities. Thus probing the mechanism of irony comprehension is a topic that is suitable for computational linguistics research. Furthermore, there are good reasons for interpreting irony by computer. Irony research can throw new light on computational studies of many pragmatic phenomena, and it can contribute to natural human-machine interaction (Hulstijn and Nijholt 1996).

Nevertheless, in the domains of computational linguistics and artificial intelligence, surprisingly little attention has been given to ironic uses of language, although other nonliteral language such as metaphor has been a popular topic (Fass, Hinkelman, and Martin 1991). The reason for this imbalance lies in the recently accepted argument that irony is a more complicated phenomenon than was supposed (Barbe 1995). Indeed, irony is far more than mere opposition of the literal meaning, and it does not always include surface incongruity or violation of cooperative norms that the traditional pragmatic view (e.g., Grice 1975; Searle 1979) assumes to be an essential feature of irony. Because of this complexity, all previous approaches to irony still do not give a plausible and computationally feasible answer to the essential questions: "What is irony?" and "How do people interpret irony?"

For this purpose, we have proposed *an implicit display view of irony* that overcomes several difficulties involved in previous irony theories (Utsumi 1996, 1997). The main claim of the implicit display view is twofold: (i) verbal irony presupposes *ironic environment*, a proper situational setting in the discourse context and (ii) verbal irony is viewed as an utterance that *implicitly displays* ironic environment.

On the basis of the implicit display view, this paper proposes a computational model of the cognitive mechanism for interpreting verbal irony in English.[1] The model is psychologically plausible in that it is consistent with various empirical findings on irony. An interpretation algorithm which embodies the model first judges whether a given utterance is potentially ironic by checking the utterance for implicit display, and then examines whether the discourse context meets requirements of ironic environment. In the rest of this paper, we present the implicit display view together with the weakness of previous irony theories in Section 2. Then we give an interpretation model of irony in Section 3 and its computational implementation in Section 4.

[1] It must be noted that this paper focuses only on verbal irony which should be distinguished from situational irony or irony of fate (i.e., situations which is ironic).

2. IMPLICIT DISPLAY VIEW OF IRONY

To see the intuitive idea behind the implicit display view, consider a mother's utterance (1a) given in Situation 1 and the same utterance given in Situation 2.

Situation 1. A mother asked her son to clean up his messy room, but he was lost in a comic book. After a while, she discovered that his room was still messy, and said to him:

Situation 2. A mother asked her son to clean up his messy room, and he did completely. After a while, she discovered that his room was clean, and said to him:

(1) a. Your room is totally clean!

Hearers who have noticed Situation 1 have no problems understanding a mother's ironic intension in (1a), but when the remark (1a) is given in Situation 2, it is no longer ironic. In terms of the implicit display view, Situation 1 is surrounded by ironic environment, but Situation 2 is not. On the other hand, even if the following remark (1b) is made by a mother in Situation 1, it is unlikely to be ironic.

(1) b. Your room is totally messy!

It merely describes the real state of affairs. In terms of the implicit display view, it does not implicitly display ironic environment although (1a) does.

Formally, ironic environment consists of the following three events:

1. the speaker has a certain expectation E.
2. the speaker's expectation E is incongruous with the reality.
3. the speaker has a negative attitude toward the incongruity (e.g., reproach, disappointment, anger, criticism).

When the discourse context satisfies these three conditions, we say that the situation is surrounded by ironic environment. For example, Situation 1 is surrounded by ironic environment since the ironist mother's expectation that his room is clean has not been fulfilled and she is disappointed with or angry at the incongruity, whereas Situation 2 is not surrounded by ironic environment because it includes no apparent incongruity between her expectation and the reality.

Implicit display of ironic environment is accomplished by an utterance U as follows:

1. U *alludes* to the speaker's expectation to the extent that coherence relations — e.g., *Volitional-Cause*, *Enable* — similar to the relations of Rhetorical Structure Theory (Mann and Thompson 1987) hold between what is said and what is expected.
2. U includes *pragmatic insincerity* by intentionally violating (or flouting) pragmatic principles — e.g., the maxim of quality, felicity conditions for speech acts, politeness principle, the maxim of quantity.
3. U indirectly expresses the speaker's negative attitude by being accompanied by a variety of *cues* — e.g., hyperbolic words and phrases (Kreuz and Roberts 1995), speech acts of "expressives", interjections, prosodic cues like so-called ironic tone of voice.

In the example above, the utterance (1a) satisfies all these conditions and thereby implicitly displays ironic environment. First, it mentions, and thus alludes to, her expectation of the room being clean. Second, it is a literally false statement that violates the maxim of quality. Third, the hyperbolic word "totally" is used to exaggerate the ironic attitude. On the

other hand, the utterance (1b) does not allude to the expectation and it is a pragmatically appropriate (i.e., sincere) utterance.

The implicit display view essentially differs from and is better than Wilson and Sperber's (1992) echoic interpretation theory, which has been the dominant view of irony.[2] The echoic interpretation theory argues that verbal irony is a variety of echoic interpretations of someone's thought, utterance, expectation or general desires / norms, in which the speaker dissociates herself from the echoed materials with ridicule or scorn. For example, Peter's echoic reply (2) of the following exchange is a typical example of echoic irony.

Situation 3. David said "I'd be promoted before you" to his colleague Peter. This elicited the following reply:

(2) Oh! you'd be promoted before me.

In the same way, the ironic utterance (1a) of Situation 1 echoes the mother's expectation with negative attitude.

The most important difference between allusion and echoic interpretation lies in what materials are echoed / alluded to by ironic utterances. On our view, ironic utterances are always motivated by, and thus, allude to the speaker's expectations, while the echoic interpretation theory argues that irony echoes not only the speaker's expectation, but also other materials. However, in the following exchange between Peter and his other colleague James, who did not know what David said,

(3) a. James: What did David said to you?
b. Peter (with ridiculing aversion): He'd be promoted before me.

Peter's utterance (3b) echoes David's preceding utterance and he simultaneously dissociates himself from the David's opinion echoed in the same way as (2), but no irony results. Hence, the echoic interpretation theory cannot distinguish irony from nonironic echoes completely. On the other hand, the implicit display view provides a consistent explanation of Peter's utterances (2) and (3b). In the case of Situation 3, what really makes the utterance (2) ironic is not David's preceding utterance itself, but the speaker Peter's expectation that the addressee David should know that his opinion expressed by the preceding utterance is false.[3] The reason that (3b) is not ironic is that the addressee James does not (and cannot) assume any irony-motivating expectation of the speaker relevant to the current exchange, and thereby, the discourse situation of (3b) is not surrounded by ironic environment.

Another point that differentiates allusion of the implicit display view from echoic interpretation is what relations are allowed between an ironic utterance and its echoed / alluded expectation. According to Wilson and Sperber (1992), an utterance U is an echoic interpretation of another utterance or an expectation to the extent that these two propositions share logical and contextual implications (i.e., resembles each other). On the other hand, the implicit display view argues that the relation between an utterance U and an alluded expectation E is best analyzed by coherence relations: Given the propositional content P of U or P's constituents P_i (we assume $P_0 = P$), and the speaker's expected event / state Q, the utterance U *alludes* to the speaker's expectation E if and only if there is a path that coherently relates P_i to Q (and U does not directly express E). For example, each of the

[2]The implicit display view copes with many problems posed by other previous irony theories. For further details, see (Utsumi 1996, 1997).

[3]Kaufer (1981) takes a similar view of irony. He argued that, in order for clearly false utterances like "Columbus discovered America in 1900" to be perceived as irony, such utterance must be given in the contextual setting in which "the ironist knows the utterance is false (and thus rejects it), knows that the addressee does not know this, and (most importantly) also *believes that the latter should know it* (*ibid.*, p.503; italics are added by the author)."

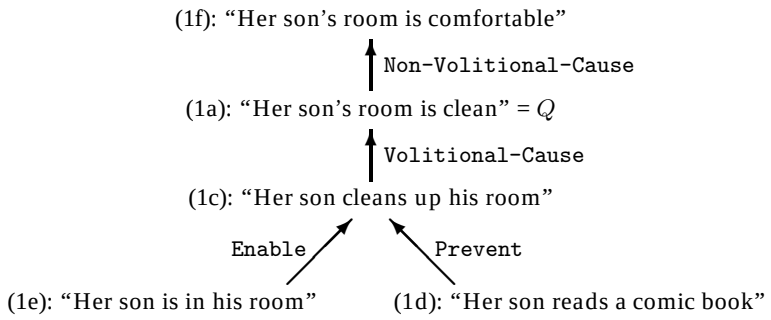

FIGURE 1. Allusion structure for Situation 1

ironic utterances (1a), (1c)∼(1f) in Situation 1 alludes to the mother's expectation as shown in Figure 1.

(1) c. I love children who keep their rooms clean.
d. Thank you for reading a comic book.
e. Oh, where is my son?
f. How do you feel in the comfortable place?

However, the echoic interpretation theory cannot account for some ironies such as (1d) and (1e) because they do not share any implications with the mother's expectation. (For example, "reading a comic book" implies "not cleaning up the room," which then implies "a messy room," but such implication is the opposite of the expectation "a clean room.")

3. IRONY INTERPRETATION MODEL

According to the implicit display view, irony is distinguished from nonirony in accordance with the two conditions: irony is given in the situation surrounded by ironic environment (ironic environment condition) and irony implicitly displays ironic environment (implicit display condition). In other words, irony interpretation is to know that the discourse situation is surrounded by ironic environment by judging an utterance to be ironic. In order to build an interpretation model of irony, however, several empirical findings and suggestions described below must be taken into account.

(I) It is unlikely to assume that people consider ironic environment whenever they interpret utterances in ordinary verbal communication because most utterances are nonironic. Thus an interpretation model of irony must screen out clearly nonironic utterances first, and consider ironic environment only after judging an utterance to be potentially ironic.

(II) Irony can be communicated even when hearers do not notice ironic environment beforehand, and thus, they cannot recognize to satisfy all the components for implicit display. For example, people can perceive the remark (4) as ironic even though they are unaware of the events of her morning and / or her expectation and they recognize neither pragmatic insincerity nor allusion to the expectation (Barbe 1995).

Situation 4. You see your friend at work for the first time that day, and she says:

(4) I've had a great morning!

Gibbs and O'Brien (1991) also pointed out that irony can be interpreted without ironic cues. These examples indicate that people do not have to recognize all the components for ironic environment and for implicit display beforehand to interpret irony. Rather, it is more appropriate to think that when hearers (and an interpretation model) do not recognize ironic environment beforehand, but when they judge that the implicit display condition is achieved to a certain degree, they decide whether the ironic environment condition is satisfied by inferring the unrecognized components from the information provided by an utterance.

(III) The speaker's expectation is the most essential component, because the other two components for ironic environment and allusion cannot be identified unless the speaker's expectation is known.[4] Therefore, an interpretation model must deal with irony differently according as the speaker's expectation which motivates irony is known beforehand or unknown.

(IV) Positive utterances are, in general, recognized to be more ironic than negative utterances (Kreuz and Glucksberg 1989; Kumon-Nakamura, Glucksberg, and Brown 1995). Such polarity effect is considerable when the speaker's expectation is implicit, but when the expectation is explicit, there is no polarity effect (i.e., negative utterances such as "New York subways are dirty" uttered in a clean train can convey irony as appropriately as positive ones) (Kreuz and Glucksberg 1989). An interpretation model must reflect such asymmetry of irony.

Irony interpretation can be modeled as a process shown in Figure 2. In accordance with (I), many nonironic utterances are screened out by the implicit display condition without ironic environment being considered (Step 1). When an utterance satisfies the implicit display condition, the ironic environment condition is examined differently as the speaker's expectation is known or unknown, which is suggested by (III). If hearers readily recognize the allusion to the speaker's expectation they know, they only examine the incongruity of the known expectation and the negative attitude (Step 3). On the other hand, when hearers do not know the speaker's expectation, or do not recognize any allusion to the known expectation, the expectation must be inferred from the utterance and the discourse context (Step 2) as suggested by (II), and it is checked for whether it is incongruous with the situation and whether a negative attitude can be elicited (Step 3). Then, in both cases, if hearers are successful in recognizing ironic environment, they judge the utterance as ironic and become aware of the speaker's ironic intention of drawing hearers' attention to and conveying the fact that the three components for ironic environment hold in the current situation.

The implicit display condition can be defined as the following formula so that it is consistent with the empirical findings (II)-(IV) and it can be used for an indicator of being potentially ironic:

$$d(U) = \begin{cases} d_A + d_I + d_E & \text{(if the speaker's expectation that motivates irony is known beforehand)} \\ d_D + d_I + d_E & \text{(otherwise)} \end{cases} \quad (5)$$

[4] Happé's (1993) empirical finding on autistic people's understanding of figurative language serves as empirical evidence for the importance of the speaker's expectation in irony interpretation. Autistic people generally suffer from a severe impairment in the ability to comprehend another person's belief (e.g., speaker's expectation), and Happé (1993) showed that autistic people could not understand irony although they understood metaphors correctly.

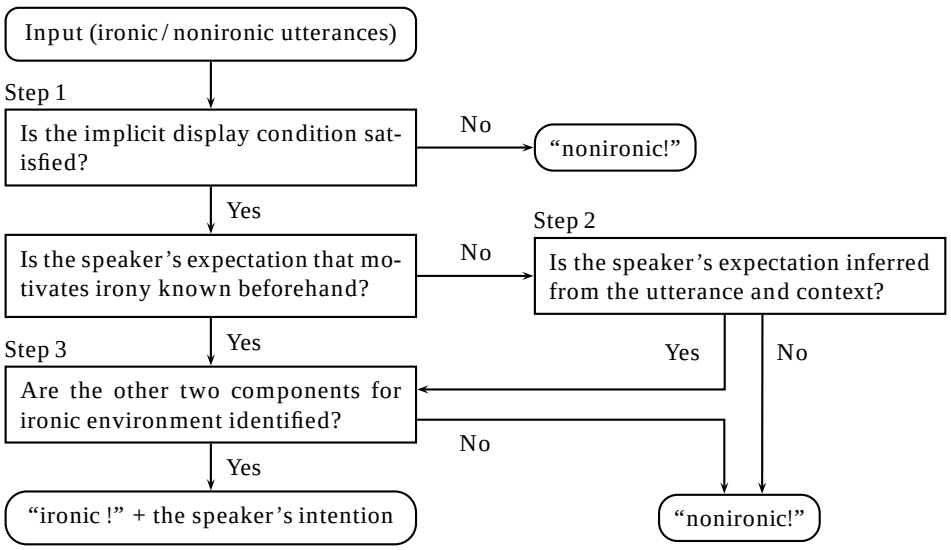

FIGURE 2. Irony interpretation model

In the formula, $d(U)$ denotes the degree of implicit display for an utterance U, d_A denotes the degree of allusion, d_I denotes the degree of pragmatic insincerity, d_E denotes the degree of indirect expression of the attitude, and d_D denotes the degree of desirability of the content of U. The formula (5) means that when the speaker's expectation that motivates irony is not known beforehand, sentence polarity (d_D) is used for a subcondition of the implicit display condition instead of allusion of the utterance (d_A), and positive utterances can facilitate ironic interpretation. On the other hand, sentence polarity does not affect the degree of implicit display when the expectation is known beforehand. Therefore, it is consistent with (III) and (IV). It also means that the implicit display condition is satisfied to the extent that its degree $d(U)$ is high, as suggested by (II). Given a certain threshold value C, an utterance does not satisfy the implicit display condition if a value of $d(U)$ is less than C. If every subcondition is either satisfied or not satisfied (e.g., d_A, d_I, d_E, d_D take either 0 or 1), it is reasonable to assume that recognition of at least two of the three components for implicit display is enough for satisfaction of the implicit display condition (i.e., $C = 2$).

Empirical evidence to support the "2-of-3" criterion for implicit display is provided by the evaluation we conducted (Utsumi 1999). In the evaluation, after reading 48 utterances with paragraph-length contexts which can be interpreted ironically, 48 subjects (graduate students) were asked to write down the speaker's expectation, and to rate the degree of ironicalness and all the components for implicit display and ironic environment on 7-point scales (0–6). The result was that utterances judged to satisfy the "2-of-3" criterion were rated as significantly more ironic than utterances judged not to satisfy, but there was no such difference between the group of utterances judged to satisfy all the three components and the group of other utterances.

Gibbs's (1986) time-course study of irony can be seen as additional support for the proposed model. He demonstrated that subjects significantly took less time to understand ironic remarks in the explicit contexts (i.e., contexts that contained the statements motivating an explicit echoic mention of some expectation) than to understand the same remarks in the implicit contexts (i.e., contexts that contained no such statement). This finding can be

```
 1.  (!= (MBH ?T) (mother x y))
 2.  (!= (MBH ?T) (son y x))
 3.  (!= (MBH ?T) (room a))
 4.  (!= (MBH ?T) (owns y a))
 5.  (!= (MBH ?T) (comic-book b))
 6.  (!= (MBH ?T) T1>T0)
 7.  (!= (MBH T0) (in x a))
 8.  (!= (MBH T0) (ask x y clean-up(y,a)))
 9.  (!= (MBH T0) (do read(y,b)))
10.  (!= (MBH T1) (not (do clean-up(y,a))))
11.  (!= (MBH T1) (say x y (!= (?* T1) (clean a))))
12.  (!= (H T0) (messy a))
13.  (!= (H T0) (hope x (!= (?* ?T:?T>T0) (clean a))))
14.  (!= (H T1) (blameworthy not(clean-up(x,a)) x))
15.  (!= (H T1) (blameworthy read(y,b) x))
```

FIGURE 3. A sample discourse context for Situation 1

explained by the proposed model. In the explicit contexts, the speaker's expectations are quite manifest to the subjects, but in the implicit contexts they are not known beforehand. Thus the additional process of inferring the speaker's expectation (i.e., Step 2) is required in the implicit contexts, and as a result, the ironic utterances in the implicit contexts take longer to process.

4. COMPUTATIONAL IMPLEMENTATION OF THE MODEL

In this section, we present an irony interpretation algorithm which has been implemented in Common Lisp. The algorithm embodies the interpretation model of irony proposed in Section 3, but it is not a full-fledged implementation of the model. For example, the algorithm does not deal with natural language texts directly; some internal representations of an utterance and of its contextual information are inputted into the algorithm. Also, the ability of the algorithm to recognize the three components for implicit display is limited; coherence relations are given by hand, and not all pragmatic principles are dealt with by the algorithm. However, it must be noted that the purpose of this paper is to provide a cognitive model of irony interpretation enough to be formalized in a computable fashion; we are not concerned here with an automatic method for interpreting irony which occurs in natural language texts.

4.1. Inputs and Shared Knowledge

The interpretation algorithm plays the role of the hearer. The inputs to interpretation consist of

1. the discourse context W, which is the set of hearer's beliefs about events / states represented by *formulas* as shown in Figure 3;
2. the propositional content P of a given utterance U represented by formulas which are not ascribed to anyone, or by *predicates* (actions);
3. the literal (surface) illocutionary act F of U, one of seven act types *Inform, Ask-if, Ask-ref, Request, Offer, Thank, Apologize*;

Causal Relations:

```
16.  (=> (!= (?B ?T0) (and (in ?X ?A) (free ?X)))
         clean-up(?X ?A)
         (!= (?B ?T1) (clean ?A)))
17.  (=> (!= (?B ?T) (do read(?X ?A)))
         (!= (?B ?T) (not (free ?X))))
18.  (=> (!= (?B ?T) (messy ?A))
         (!= (?B ?T) (not (clean ?A))))
19.  (=> (!= (?B ?T) (clean ?A))
         (!= (?B ?T) (comfortable ?A)))
20.  (<=> (!= (?B ?T) (blameworthy ?A))
          (!= (?B ?T) (not (praiseworthy ?A))))
```

Speech Act Schemes:

```
21.  (=> (ascribe ?P ?S)
         Inform(?S,?H,?P)
         (!= (H ?T1) (intend ?S Convince(?S,?H,?P))))
22.  (=> (!= (SH ?T0) (and (do ?P) (praiseworthy ?P ?S)))
         Thank(?S ?H,?P)
         (!= (H ?T1) (intend ?S Convince(?S,?H,(!= (SH ?T0) (grateful ?S ?H ?P))))))
```

Emotion-Eliciting Rules:

```
23.  (<=> (!= (?B ?T) (hope ?X ?I))
          (!= (?B ?T) (and (want ?X ?I) (expect ?X ?I))))
24.  (=> (and (!= (?B ?T0) (hope ?X (!= (?B1 ?T:?T>?T0) (not ?I))))
              (!= (?B ?T1:?T1>?T0) ?I))
         (!= (?B ?T1) (disappointed ?X (!= (?B1 ?T) ?I))))
25.  (=> (and (!= (?B ?T0) (hope ?X (!= (?B1 ?T:?T>?T0) (not ?I))))
              (!= (?B1 ?T1:?T1>?T0) (and ?I (do ?A) (blameworthy ?A ?X)))
              (vol-cause ?A (!= (?B1 ?T1) ?I)))
         (!= (?B ?T1) (angry-at ?X Agent(?A) ?A)))
26.  (=> (!= (?B ?T) (and (do ?A) (blameworthy ?A ?X)))
         (!= (?B ?T) (reproach ?X Agent(?A) ?A)))
```

FIGURE 4. An example of the shared knowledge used in the system

4. the feature-based semantic representation M of U (e.g., `(Rel:sem1, Theme:M1)`); and
5. the set of the shared knowledge K consisting of *causal relations* as domain knowledge, *speech act schemes* and *emotion-eliciting rules*, some of which are shown in Figure 4.

We assume that the three inputs 2–4 can be identified by the parser.[5]

The irony interpreter uses the situation-theoretic representation scheme (Utsumi 1996). All events and states are expressed as formulas `F=(!= (B T) I)`, support relations between *situations* `(B T)` and *contents* `I`. This formula is identical to the situation-theoretic notation $(B,T) \models I$ in which a situation (B,T) supports an infon I (i.e., (B,T) makes I true). A situation consists of *a belief space* `B` and *time* `T` of the event/state. As a belief space, we

[5]Prosodic and nonverbal features must also be taken into account for spoken ironic language, but computational formalization of these features is beyond the scope of this paper.

use "H" (hearer's beliefs), "SH" (hearer's beliefs about speaker's beliefs), and "MBH"(hearer's one-sided mutual beliefs (Clark and Marshall 1981)). The formula (neg F) denotes (!= (B T) (not I)) and ¬F means that (B T) does not support I. Symbols prefixed with "?" are universally quantified variables, and variables with ":" are restricted ones. For example, Formula 12 (!= (H T0) (messy a)) in Figure 3 expresses the state that the hearer believes the son's room is messy at t_0, and the formula (!= (?* T1) (clean a)) is the propositional content of the utterance (1a). Note that since the content of (1a) is not ascribed to anyone, its belief space is represented by a variable "?*". Propositional contents are also represented by predicates: for example, the action read(y,b) is the propositional content of (1d).

A causal relation between two events/states is expressed by (=> (!= Sit1 I1) A (!= Sit2 I2)). This relation means that if an action A is executed in Sit1 supporting I1, then it causes I2 in the resulting situation Sit2. A non-volitional causal relation is also represented as (=> (!= Sit1 I1) (!= Sit2 I2)). Coherence relations for allusion are defined using these causal relations. For example, given that (=> F1 A F2) and (=> F3 F4), it follows that (vol-cause A F2), (enable F1 A), (non-vol-cause F3 F4), and (prevent (neg F1) A). Furthermore, from (=> F5 B (neg F1)) it follows that (prevent B A).

Speech act schemes are also represented as causal relations. In the schemes of Figure 4, ?S, ?H and ?P denote the speaker, the hearer and the propositional content of the utterance, respectively. Furthemore, (ascribe ?P ?X) denotes the formula generated by ascription of ?P to ?X, and it expresses an agent ?X's belief that ?P is true. For example, the ascribed content of (1a) to the speaker ?S = x is (ascribe (!= (?* T1) (clean a)) x) = (!= (SH T1) (clean a)).[6]

Emotion-eliciting rules we use are originally proposed by O'Rorke and Ortony (1994), but differ in that our rules explicitly distinguish the situation in which a person feels an emotion from the situation of the events/states toward which he/she feels an emotion. In this paper, we assume that "expect" and "want" are primitive emotions for representing the speaker's expectation, and the emotion of "hope" is a compound of "expect" and "want" (as shown by Formula 23 of Figure 4). We limit the speaker's negative emotions to "disappointment", "anger" and "reproach" (as shown by Formulas 24-26).

4.2. Algorithm

Figure 5 shows an interpretation algorithm which embodies the interpretation model of irony in Figure 2.[7] The algorithm answers whether an inputted utterance U is ironic, and when U is ironic it returns the speaker's ironic intention. In Figure 5, C denotes the union set of W and K, and (ask F C) is a function that answers whether a query F is entailed by C and returns one possible substitution that makes the query true. In the entailment, we use the following inference rules: "if (!= (?B ?T0) ?I) and ¬(!= (?B ?T:?T>?T0) (not ?I)), then (!= (?B ?T) ?I)" (frame axiom) and "if (!= (H T) I) and ¬(!= (SH T) (not I)) then (!= (SH T) I)" (default ascription).

As an example, let us consier how the algorithm interprets (1a) in Situation 1, assuming that W is the context of Figure 3.

Situation 1. A mother asked her son to clean up his messy room, but he was lost in a comic book. After a while, she discovered that his room is still messy, and said to him:

[6] Note that this formula expresses the speaker's belief from the point of view of the hearer. Hence, from the the point of view of the speaker, the same belief is expressed by (!= (S T1) (clean a))

[7] Although this algorithm assumes that d_A, d_I, d_E, d_D take binary values, it can be easily modified for quantitative measurement. For example, we can caluculate the degree of allusion by $d_A = 1 - 0.1n$ in which n denotes the depth of the coherence path for allusion.

Step 1

1-1. If `(ask P C)` returns yes, the implicit display condition is not satisfied (i.e., U *directly* expresses ironic environment).

1-2. Select the speaker's expectation E = `(!= (?B ?T) (Re Xs Q))` from W such that `Re` $\in$ {`expect`, `want`, `hope`}, `Xs` denotes the speaker, and `Q` is the context of the expectation.

1-3. If E is found, for $P_i \in$ {the set of all constituent formulas and actions of P}, find a path that coherently relates `Q` to P_i by a breadth-first search. Note that the depth of a search tree is limited to 5.

1-4. If E is not found or no allusion to E is found, assess the polarity of U using a semantic representation M.

1-5. Find pragmatic insincerity of U as follows:

(a) One of the instantiated preconditions `F` of the speech act scheme of F is violated if `(ask (neg F) C)` returns yes.

(b) Check the utterance U for empirically observable patterns indicating overpoliteness, or understatements (i.e., violation of the maxim of quantity).

1-6. Find cues for implicit display of the speaker's negative attitude from M.

1-7. Calculate the degree of implicit display $d(U)$ by the formula (5). If $d(U) \geq C(= 2)$, the implicit display condition is satisfied.

Step 2 Infer E by searching a desirable (positive) event / state related to P by coherence relations.

Step 3 If the answer of (a) is yes, incongruity of the expectation is not identified. If all the answers of (b) are no, any negative attitude is not identified.

(a) Perform `(ask Q C)`.

(b) Perform `(ask (!= (?B ?T:?T>=Te) (angry-at Xs ?Xa ?A)) C)`,
`(ask (!= (MBH ?T:?T>=Te) (disappointed Xs (neg Q))) C)`, and
`(ask (!= (?B ?T:?T>=Te) (reproach Xs ?Xa ?A)) C)`.

Output Output the speaker's ironic intention, and return "ironic".

FIGURE 5. Irony interpretation algorithm

(1) a. Your room is totally clean!

The algorithm selects Formula 13 `(!= (H T0) (hope x (!= (?* ?T:?T>T0) (clean a))))` in W as the speaker's expectation E at Step 1-2, and finds that the utterance (1a) alludes to E at Step 1-3 since its content P=`(!= (?* T1) (clean a))` is unifiable to Q=`(!= (?* ?T:?T>T0) (clean a))`. At Step 1-5, the algorithm recognizes the violation of the precondition `(ascribe (!= (?* T1) (clean a)) x)` for F =*Inform*, because `(neg (ascribe (!= (?* T1) (clean a)) x))= (!= (SH T1) (not (clean a)))` is derived from Formula 12 `(!= (H T0) (messy a))` in W by the inference rules. At Step 1-6, the algorithm finds the hyperbolic word "totally" in U. From these results, Step 1 judges the utterance (1a) satisfies the implicit display condition. Since the known expectation is judged to be alluded to by (1a), it is checked for incongruity with the situation at Step 3(a) and for negative attitude at Step 3(b). In this case, `(ask (!= (?* ?T:?T>T0) (clean a)) C)` returns no (i.e., her son's room is not clean), and `(ask (!= (?B ?T:?T>=T0) (angry-at x ?Xa ?A)) C)` returns yes (i.e., the speaker's negative emotion is elicited using the emotion-eliciting rule for anger, i.e.,

Formula 25 of Figure 4). As a result, the algorithm judges (1a) to be ironic, and produces the following intention:

```
Inform(x,y,(and (!= (SH T0) (hope x (!= (?* ?T:?T>T0) (clean a))))
                (!= (SH T1) (not (clean a)))
                (!= (SH T1) (angry-at x y not(clean-up(y,a))))))
```

In the same way, the algorithm judges other ironies (1c)∼(1f) to be ironic. For example, the utterance (1d) is judged to achieve implicit display at Step 1: its propositional content `read(y,b)` is found to be coherently related to Q=`(!= (?* ?T:?T>T0) (clean a))` by the following path at Step 1-3,

```
read(y,b) --- prevent ---> clean-up(y,a)              (by Formulas 16 and 17)
          --- volitional-cause ---> (!= (?B ?T) (clean a))   (by Formula 16)
          --- unifiable ---> (!= (?* ?T:?T>T0) (clean a))
```

and the violation of the precondition `(!= (SH ?T0) (praiseworthy read(y,b) x))` for the illocutionary act *Thank* is recognized at Step 1-5 since `(!= (SH ?T0) (not (praiseworthy read(y,b) x)))` is derived from C.

On the other hand, if W does not include Formula 13 `(!= (H T0) (hope ... (clean a)))`, Step 2 infers the speaker's expectation from the content of (1a). Since "being clean" is judged to be positive at Step 1-4, the algorithm derives an assumption that the speaker expects that the room is clean, and Step 3 checks it for the other components. Note that when a given utterance is not positive, a positive proposition is generated using coherence relations by the same search method as Step 1-3.

The algorithm also rejects nonironic utterances correctly. For example, the utterance (1b) (i.e., "Your room is totally messy!") in Situation 1 is rejected at Step 1-1, because `(ask (!= (?* T1) (messy a)) C)` returns yes. Similarly, the utterance (1a) in Situation 2 is judged to be nonironic at Step 3(a) because the query `(ask (!= (?* ?T:?T>T0) (clean a)) C)` answers yes (i.e., the expectation has been realized), although it is judged to satisfy the implicit display condition.

4.3. Related Work

The model we have proposed here is the first computational implementation of irony interpretation, but one notable computational study is made on automatic detection of Japanese irony by Takizawa and Ito (1994). Although their study is not intended as an investigation of the cognitive mechanism of irony interpretatin, it may be worth discussing the relation and the difference between our model and Takizawa and Ito's (1994) model.

Takizawa and Ito's (1994) algorithm for detecting irony takes as input a frame-based representation of an utterance and of a situation, and then caclulates the degree of ironicalness as the product of the following three numerical measures.

1. The degree of incompatibility between the utterance and the situation
2. The strength of the causal relationship between the utterance and the situation
3. The presence of ironic markers (i.e., terminating particles in Japanese)

These measures correspond to the three components for implicit display. The degree of incompatibility is assumed to take a high value when the utterance is positive but the situation is negative, and thereby, it can be seen as a limited measure of pragmatic insincerity (and sentence polarity). Likewise, the strength of the causal relationship can be seen as measuring our allusion in part, and ironic markers is subsumed under the cues for indirect

expression of the speaker's negative attitude. Hence, Takizawa and Ito's method partially calculates to what degree an utterance satisfies the implicit display condition. At the same time, however, their method cannot deal with the ironic environment condition, the essential condition for an utterance to be ironic: the situation inputted to their algorithm is limited to an event / state which is incongruous with the utterance (e.g., "the room is messy" or "her son does not clean up the room" in the case of Situation 1).

5. CONCLUDING REMARKS

This paper has presented our cognitive model of irony interpretation and its implementation, and has shown the validity of the model by describing how the model explains empirical findings on irony. In order to improve the model, however, we have to consider 1) perlocutionary communication goals conveyed by irony; 2) evaluation of the performance by comparing with human interpretation; and 3) prosodic and nonverbal features of irony. We are extending the model considering these issues.

ACKNOWLEDGMENTS

This research is supported in part by a grant from the Okawa Foundation for Information and Telecommunications.

REFERENCES

BARBE, K. 1995. Irony in Context. John Benjamins Publishing Company.

CLARK, H., and C. MARSHALL. 1981. Definite reference and mutual knowledge. *In* Elements of Discourse Understanding. *Edited by* A. Joshi, B. Webber, and I. Sag. Cambridge University Press, Cambridge. pp.10–63.

FASS, D., E. HINKELMAN, and J. MARTIN. 1991. Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idioms, Speech Acts, Implicature.

GIBBS, R. 1986. On the psycholinguistics of sarcasm. Journal of Experimental Psychology: General, **115**: 3–15.

GIBBS, R., and J. O'BRIEN. 1991. Psychological aspects of irony understanding. Journal of Pragmatics, **16**: 523–530.

GRICE, H. 1975. Logic and conversation. *In* Syntax and Semantics, Vol.3: Speech Acts. *Edited by* P. Cole, and J. Morgan. Academic Press, New York. pp.41–58.

HAPPÉ, F.G.E. 1993. Communicative competence and theory of mind in autism: A test of relevance theory. Cognition, **48**: 101–119.

HULSTIJN, J., and A. NIJHOLT. 1996. Proceedings of the International Workshop on Computational Humor.

KAUFER, D. 1981. Understanding ironic communication. Journal of Pragmatics, **5**: 495–510.

KREUZ, R., and S. GLUCKSBERG. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. Journal of Experimental Psychology: General, **118**(4): 374–386.

KREUZ, R., and R. ROBERTS. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. Metaphor and Symbolic Activity, **10**(1): 21–31.

KUMON-NAKAMURA, S., S. GLUCKSBERG, and M. BROWN. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. Journal of Experimental

Psychology: General, **124**(1): 3–21.

MANN, W., and S. THOMPSON. 1987. Rhetorical structure theory: Toward a functional theory of text organization. Text, **8**(3): 167–182.

O'RORKE, P., and A. ORTONY. 1994. Explaining emotions. Cognitive Science, **18**: 283–323.

SEARLE, J. 1979. Expression and Meaning. Cambridge University Press, Cambridge.

TAKIZAWA, O., and A. ITO. 1994. A method for detecting an ironic expression. Journal of Japanese Society of Artificial Intelligence, **9**(6): 875–881, in Japanese. (For an brief English description of the method, see (Takizawa, Yanagida, Ito, and Isahara 1996)).

TAKIZAWA, O., M. YANAGIDA, A. ITO, and H. ISAHARA. 1996. On computational processing of rhetorical expressions: Puns, ironies and tautologies. *In* Proceedings of the International Workshop on Computational Humor (IWCH'96). pp.39–52.

UTSUMI, A. 1996. A unified theory of irony and its computational formalization. *In* Proceedings of the 16th International Conference on Computational Linguistics (COLING 96). pp.962–967.

UTSUMI, A. 1997. What's irony?: Implicit display theory of verbal irony. Cognitive Studies, **4**(4): 99–112, in Japanese.

UTSUMI, A. 1999. How is irony distinguished from nonirony?: An implicit-display-based model of irony-nonirony distinction. Journal of Japanese Society of Artificial Intelligence, **14**(4), in Japanese.

WILSON, D., and D. SPERBER. 1992. On verbal irony. Lingua, **87**: 53–76.