An Analysis of Metareference

SETH EBNER
Johns Hopkins University
December 22, 2017

Introduction

To anyone who has watched a film, it is clear that there are many different contexts in which entities exist. Hogwarts is not a real place. Luke Skywalker is not a real person, nor is his lightsaber a real-life weapon. Specifically, for a given piece of fictional material, there are two contexts (models) at play: the "real-world" context inhabited by the audience, and the "fictive" context inhabited by the fictional characters. In many cases, these two contexts never interact. However, that isn't to say that they cannot interact. One such way they interact is through metareference. Metareference is "a special, transmedial form of [...] self-reference [...] within an artefact or performance" [6]. That is to say, metareference involves some piece of the fictional work demonstrating awareness that it is in such a fictional work. Is there a way to provide a formal semantic account of when metareference occurs?

Metareference and Markedness

Metareference is a form of self-awareness in which fictional characters express the notion that they know that they are in a film, novel, etc. These expressions are marked and often evoke a humorous response from the audience. Metareference is a technique related to breaking the fourth wall, in which characters address the audience directly.

For example, in *Spaceballs*, the character Dark Helmet watches a screen playing the movie *Spaceballs* itself and ponders his next move in light of the realization that he is in a movie. *Monty Python and the Holy Grail* contains many metareferential moments, including ones in which the Arthurian characters are arrested by modern police and in which they charge through the filming of a documentary.

A Puzzle

It is seemingly easy for humans to notice metareference, but it is not always quite as easy to pinpoint why a scene is metareferential. We seek to develop a set of criteria that can consistently detect metareference.

We motivate our problem with a puzzle. Consider the following dialogues, adapted from the film *Deadpool*:

(1)

COLOSSUS: Let us go talk to the Professor.

DEADPOOL: Xavier?

(2)

COLOSSUS: Let us go talk to the Professor.

DEADPOOL: McAvoy?

Dialogue (1) is not unusual in any way. Its referents stay completely within the universe of the film. As a result, (1) is an unmarked discourse, and a viewer of the film wouldn't give it second thought.

In contrast, the referent "McAvoy" in dialogue (2) is not part of the film's universe; it is an entity only in the "real world". "McAvoy" refers to an actor who portrayed Professor Xavier in previous films. Dialogue (2) is marked due to this mixing of contexts, and it leads to a humorous moment. Specifically, the reference to the actor "McAvoy" (as opposed to some other "real world" reference) makes (2) an example of metareference.

The only difference between (1) and (2) is the referent in Deadpool's response. Colossus's utterance and Deadpool's tone are constant. *So what causes the metareference to occur?*

The Data

In the following examples, words triggering metareference are noted in blue. The examples have been adapted from their source for ease of analysis. Those marked with an asterisk (*) do not occur in the source material and are intended to contrast with the originals. A hash (#) indicates oddness.

(3) [**Deadpool*, repeated here for reference]

COLOSSUS: Let us go talk to the Professor.

DEADPOOL: Xavier?

(4) [*Deadpool*, repeated here for reference]

COLOSSUS: Let us go talk to the Professor.

DEADPOOL: McAvoy?

(5) [*_Deadpool_]

COLOSSUS: Let us go talk to the Professor.

DEADPOOL: #Feynman?

(6) [_The Wizard of Oz_]

DOROTHY: Toto, I've a feeling we're not in Kansas anymore.

(7) [*_The Wizard of Oz_]

DOROTHY: Toto, I've a feeling we're not on that film set anymore.

(8) [_The Disaster Artist_]

SANDY: Well, why don't we just shoot in the real alleyway?

TOMMY: Because it's a real Hollywood movie.

We note that is it difficult to devise an example of metareference within _The Disaster Artist_ because it is itself a movie about filmmaking (more specifically, the making of _The Room_). This inability to create a natural example is data in itself.

Patterns in the Data

Before devising criteria for determining whether an utterance is metareferential, we take note of some general patterns in the data that may help guide our analysis.

The first point we notice is that the examples invoking metareference (4, 7) seem to require some extra cognitive effort to interpret in context, in contrast to those examples that are unmarked (3, 6, 8). However, (4) and (5) (which is not an instance of metareference) both share this characteristic extra effort, so we stipulate a unidirectional implication: metareference materially implies extra cognitive effort. This does not provide direct support for labeling an utterance as metareferential. However, the contrapositive does provide support for labeling an utterance as not metareferential: no extra cognitive effort materially implies no metareference.

We also observe that the metareference examples do not seem to be interpretable fully within the context of the fictive universe. For example, there is no salient entity mcavoy in the _Deadpool_ universe. Furthermore, the words that trigger the metareference appear to be instantiated in the real world. mcavoy is an entity in the real world. The interpretations derived in these examples rely on information from the real world context in addition to information from the fictive context. We cannot always get an interpretation under $M_{FICTIVE}$, and we never get one under just $M_{FICTIVE}$ when metareference occurs.

Given the above observations, we would expect both (4) and (5) to give rise to metareference. To disentangle these examples, we additionally note that the metareference triggers concern some aspect of the fictive material as it is instantiated in the real world (e.g., actors, film sets, a movie's existence).

Analysis

Formalization

Before giving our analysis of the causes of metareference, we formalize the observations made from the data.

A **model** is a tuple $\langle D, I, W \rangle$, where D is the domain of individuals (a set of entities), I is an interpretation function that assigns a denotation to every constant, and W is a set of worlds.

For every context involving a fictive element, we define two models, $M_{FICTIVE}$ and $M_{REAL-WORLD}$, to be the models under which fictive world and the real world are interpreted, respectively. Specifically, $M_{FICTIVE} = \langle D_f, I_f, W_f \rangle$ and $M_{REAL-WORLD} = \langle D_{rw}, I_{rw}, W_{rw} \rangle$. Each fictive context has its own model, e.g., $M_{DEADPOOL}$ or M_{OZ}.

The set of worlds in $M_{FICTIVE}$ behaves analogously to the set of worlds in $M_{REAL-WORLD}$ (each world is how things could possibly be based on the entities in $D_{fictive}$, etc.), but they have no overlap with the worlds in $M_{REAL-WORLD}$.

We define an **interpretation** of an utterance under a given model to be the entity that is picked out by the referring expression using the resources available in that model (D, I, and W).[1] An interpretation is **available** if there is such an entity that is picked out, and it is unavailable otherwise. Two interpretations are **equivalent** if they pick out the same entity.

The Conditions

We propose that metareference occurs in an utterance in a fictional work when all of the following conditions hold:

1. There is no available interpretation of the utterance under just $M_{FICTIVE}$
2. Subsequently invoking $M_{REAL-WORLD}$ (in possible conjunction with $M_{FICTIVE}$) allows for an available interpretation of the utterance

[1] In this paper, we concern ourselves only with metareference as it pertains to entities (that is, nouns or, more generally, determiner phrases). Other syntactic categories do not seem to easily trigger metareference.

3. The interpretation of the utterance involves a term referring to (some aspect of) the film, novel, etc. itself or its creation

We may (at least as a first pass) imagine the existence of a list of triggering terms whose presence satisfies condition 3.

We take the composition of two models to be the model $\langle D_f \cup D_{rw}, I_f \cup I_{rw}, W_f \rangle$.

We assume the following flow of computations on the part of the audience: interpret the literal string in $M_{FICTIVE}$; if there is no available interpretation, interpret within $M_{REAL-WORLD}$; if there is still no available interpretation, back off to what the character could have been intending to say and interpret that in $M_{FICTIVE}$. This is somewhat similar to the forward- and backward-chaining discussed in [3].

Explanatory Power

We use M_{RW} as an alias for $M_{REAL-WORLD}$, M_{DP} as an alias for $M_{DEADPOOL}$, M_{OZ} as an alias for $M_{WIZARD-OF-OZ}$, and M_{DA} as an alias for $M_{DISASTER-ARTIST}$.

In (3), the string "the Professor" uttered by Colossus has the denotation $\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}}$, which is assigned by I_{DP} the value of xavier. That is, under M_{DP} "the Professor" is interpreted as xavier. Deadpool's response of "Xavier" has the denotation $\langle\!\langle \textbf{Xavier} \rangle\!\rangle^{M_{DP},g,w_{DP}}$, which is assigned the value of xavier by I_{DP} and the lexicon. There is an available interpretation of Deadpool's utterance, so the exchange is unmarked. Also note that the interpretation of Colossus's utterance is equivalent to the interpretation of Deadpool's utterance. Furthermore, the interpretation of Deadpool's utterance is available under just M_{DP} , so there is no metareference.

In (4), the string "the Professor" uttered by Colossus again is interpreted as xavier. Deadpool's response of "McAvoy" has the denotation $\langle\!\langle \textbf{McAvoy} \rangle\!\rangle^{M_{DP},g,w_{DP}}$, which is assigned the value of $\#_e$ by I_{DP} and the lexicon because there is no such entity referred to as "McAvoy" under M_{DP} .

We now consider Deadpool's interpretation of Colossus's utterance and how it could have led Deadpool to equate "McAvoy" with "the Professor". Because it would be infelicitous and cause a "Hey, wait a minute" reaction for Deadpool to refer to an entity that does not exist (or he believes to not exist), we assume the existence of some entity mcavoy in some model. That is, reference to an entity presupposes that entity's existence. $\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}} = \mathrm{xavier} \neq \mathrm{mcavoy}$, or more generally, $\mathrm{mcavoy} \notin \langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}}$. There is no available interpretation under M_{DP} (condition 1).

We now interpret "the Professor" in $M_{REAL-WORLD}$ and w_{RW}. $\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{RW},g,w_{RW}}$ = {$\mathrm{feynman}$, $\mathrm{hawking}$}, of which mcavoy is not an element. There is no interpretation of "the Professor" in M_{RW} that yields mcavoy.

We back off again, this time to $\langle\!\langle \textbf{actor(the Professor)} \rangle\!\rangle^{M_{DP},g,w_{DP}}$. This yields

$$\langle\!\langle \textbf{actor(the Professor)} \rangle\!\rangle^{M_{DP},g,w_{DP}}$$

$= \langle\!\langle \textbf{actor} \rangle\!\rangle^{M_{DP},g,w_{DP}}(\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}})$

$= \#(\mathrm{xavier})$

$= \#_e$

Under $M_{REAL-WORLD}$:

$\langle\!\langle \textbf{actor(the Professor)} \rangle\!\rangle^{M_{RW},g,w_{RW}}$

$= \langle\!\langle \textbf{actor} \rangle\!\rangle^{M_{RW},g,w_{RW}}(\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{RW},g,w_{RW}})$

$=$ actor($\mathrm{feynman}$) OR actor($\mathrm{hawking}$)

$= \#_e$ OR $\#_e$

$= \#_e$

Combining the models:

$\langle\!\langle \textbf{actor} \rangle\!\rangle^{M_{RW},g,w_{RW}}(\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}})$

$=$ actor(xavier)

$= \mathrm{mcavoy}$

We see that Deadpool's utterance in (4) can be interpreted if we backoff to evaluating the *actor* function in $M_{REAL-WORLD}$ on the argument xavier in M_{DP} (condition 2). Furthermore, the use of the *actor* function means that this interpretation satisfies condition 3. The conditions specify that metareference has occurred, which is in line with our judgment on (4).

In (5), the interpretation of Colossus's utterance again is xavier. Deadpool's response of "Feynman" has the denotation $\langle\!\langle \textbf{Feynman} \rangle\!\rangle^{M_{DP},g,w_{DP}}$, which is assigned the value of $\#_e$ by I_{DP} and the lexicon because there is no such entity referred to as "Feynman" under M_{DP} .

As in our analysis of (4), we now consider Deadpool's interpretation of Colossus's utterance and how it could have led Deadpool to equate "Feynman" with "the Professor". We again assume the existence of some entity feynman in some model. $\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}} = \mathrm{xavier} \neq \mathrm{feynman}$, or more generally, $\mathrm{feynman} \notin \langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{DP},g,w_{DP}}$. There is no available interpretation under M_{DP} (condition 1).

We now interpret "the Professor" in $M_{REAL-WORLD}$ and w_{RW}. $\langle\!\langle \textbf{the Professor} \rangle\!\rangle^{M_{RW},g,w_{RW}} = \{\mathrm{feynman}, \mathrm{hawking}\}$, of which feynman *is* an element. There is an interpretation of "the Professor" in M_{RW} that yields feynman (condition 2).

However, no term in the interpretation satisfies condition 3, so we conclude that metareference does not occur. Also note that Colossus's utterance is interpreted entirely within M_{DP} , but Deadpool's utterance is interpreted entirely within $M_{REAL-WORLD}$, which causes Deadpool's utterance to seem odd.

Dorothy's utterance in (6) is unmarked. "Toto" picks out an entity in D_{OZ}, so there is an available interpretation (namely, toto). In (6) and (7), we focus on interpreting the determiner phrase specifying the location ("Kansas" and "film set") and seek an explanation for why (6) does not display metareference but (7) does.

The audience knows that *The Wizard of Oz* takes place in Kansas, so they accommodate an entity kansas in D_{OZ}, and $I_{OZ}(\text{Kansas}) = \mathrm{kansas}$. That is, the denotation of $\langle\!\langle \textbf{Kansas} \rangle\!\rangle^{M_{OZ},g,w_{OZ}}$ is interpreted as kansas. There is an interpretation under M_{OZ} , so there is no metareference.

In contrast, there is no notion of a "film set" in M_{OZ} : $\langle\!\langle \textbf{film set} \rangle\!\rangle^{M_{OZ},g,w_{OZ}} = \#_e$. This satisfies condition 1. There is an entity $\mathrm{film\_set}$ in D_{RW}, however, that is the film set used in *The Wizard of Oz*. (There may be many more film sets, e.g., those used for other movies, but those are different entities: $\mathrm{film\_set\_2}$, $\mathrm{film\_set\_3}$, etc.) $\langle\!\langle \textbf{film set} \rangle\!\rangle^{M_{RW},g,w_{RW}} = I(w_{RW}, \text{film set}) = \mathrm{film\_set} \in D_{RW}$. This interpretation satisfies condition 2.

Furthermore, the interpretation satisfies condition 3, as $\mathrm{film\_set}$ concerns some aspect of the film itself in M_{RW} . We expect metareference to occur in (7), and it does.

We now turn to *The Disaster Artist* (8), which is a film depicting the creation of the real-world film *The Room*. The pronoun "it" in Tommy's response refers to (the fictional recreation of) *The Room*, which exists in D_{DA}. After resolving this coreference, we have $\langle\!\langle \textbf{\textit{The Room}} \rangle\!\rangle^{M_{DA},g,w_{DA}} = \mathrm{the\_room}$. There is an interpretation under M_{DA} , so metareference does not occur.

Note that there is an interpretation under $M_{REAL-WORLD}$ (i.e., "it" refers to *The Disaster Artist*) that could trigger condition 3 (the utterance mentions the movie's being a "real Hollywood movie"), but that interpretation requires more computation on the part of the audience, as $M_{REAL-WORLD}$ would have to be composed with M_{DA} . The computationally easiest interpretation is the one given above that only concerns M_{DA} , and this interpretation does not cause condition 3 to trigger. Even though condition 3 appears on the surface to hold, conditions 1 and 2 must also be satisfied to determine whether metareference occurs.

The difficulty of creating examples of metareference in *The Disaster Artist* arises because a reference to *The Disaster Artist* is more readily applied to the in-film depiction of *The Room*. It is computationally easier for the audience to interpret an utterance within M_{DA} , so they will prefer such an interpretation (and attribute it to *The Room*) over one that invokes $M_{REAL-WORLD}$ (and attribute it to *The Disaster Artist*). This tendency to create an available interpretation means that condition 1 will not be satisfied, and so metareference will not occur.

Predictive Power

We have shown that the conditions outlined above can explain the examples presented earlier. Now we present some new data and see how the proposed analysis fares.

(9) [**Deadpool*]

COLOSSUS: Let us go talk to the Professor.

DEADPOOL: #Will Smith?

In this example, Deadpool refers to an actor, but not one who portrays any character in the film. Condition 1 is satisfied: given just M_{DP} , there is no valid interpretation of $\langle\langle$**Will Smith**$\rangle\rangle^{M_{DP},g,w_{DP}}$ (in part because there is no such entity $\mathrm{will\_smith}$ in D_{DP}). Condition 2 is also satisfied because $\langle\langle$**Will Smith**$\rangle\rangle^{M_{RW},g,w_{RW}}$ has an interpretation: $\mathrm{will\_smith} \in D_{RW}$. Condition 3, however, does not hold because $\mathrm{will\_smith}$ does not pertain to the film *Deadpool* in any way. We therefore predict no metareference in this example.

The above analysis shows that condition 3 is necessary: the criterion for metareference cannot be simply that $M_{FICTIVE}$ and $M_{REAL-WORLD}$ are both used in a valid interpretation. This point is further supported by the observation that we can evaluate predicates like *tall′(*deadpool*)*, which are nominally evaluated in $M_{FICTIVE}$ but are actually evaluated in $M_{REAL-WORLD}$ with respect to the actor. In both the previous example and this example predicate, there is no metareference, and correspondingly conditions 1 and 2 are satisfied but condition 3 is not. Thus, condition 3 is necessary.

(10) [**The Wizard of Oz*]

DOROTHY: #Toto, I've a feeling we're not in Camelot anymore.

This example is of another odd utterance. Condition 1 is not satisfied even though "Camelot" is a fictional location because it is not from the same world as *The Wizard of Oz*. So, we (correctly) predict that there is no metareference.

(11) [Hypothetical movie]

A: How can they just blow up the Empire State Building like that?

B: Because it's a real Hollywood movie.

Because the previous examples all only changed the second line of dialogue (if there is a second line), we change the first line of dialogue in (11) to result in metareference; the second line remains the same as that in (8).

There is no movie-within-a-movie plotline in the hypothetical film in (11), so "it" can only refer to the movie itself as it exists in the real world. (We assume $\mathrm{hyp\_movie} \in D_{RW}$.) The villains' blowing up of the Empire State Building is not able to be "a real Hollywood movie". $\langle\!\langle \textbf{it} \rangle\!\rangle^{M_{HYP}, g, w_{HYP}} = \langle\!\langle \textbf{\textit{Hypothetical movie}} \rangle\!\rangle^{M_{HYP}, g, w_{HYP}} = \#_e$. There is no available interpretation under M_{HYP} , so condition 1 is satisfied.

Incorporating $M_{REAL-WORLD}$ yields an available interpretation: $\langle\!\langle \textbf{it} \rangle\!\rangle^{M_{RW}, g, w_{RW}} = \langle\!\langle \textbf{\textit{Hypothetical movie}} \rangle\!\rangle^{M_{RW}, g, w_{RW}} = \mathrm{hyp\_movie}$. The referent of the pronoun "it" is the movie (and the utterance is more generally an assertion of the movie's existence or state), so condition 3 is satisfied as well. Metareference is predicted to occur, and it does.

(12) [Real-world conversation]

A: Who is the Professor in *Deadpool*?

B: I hope it's McAvoy.

The above example is an exchange between two real-world people; there is no $M_{FICTIVE}$. There is nothing to be self-aware of, so we expect there to be no metareference. Indeed, condition 1 is not satisfied because there is no $M_{FICTIVE}$ to use in an interpretation, yielding the prediction that there is no metareference.

Potential Problems

There are a couple questions still left to resolve. One issue is how the list governing condition 3 is constructed. We have assumed throughout our analysis that such a list exists, but not how it was created. It is unclear at this point how the list is created beyond hard-coding its elements.

Another issue concerns why (6) is unmarked but (5) is marked. We briefly address this in the following section.

The Pragmatics of Metareference

A brief discussion of the pragmatics of metareference is in order after having presented a semantic analysis. Again, in metareference, some component of the fictive context is aware of—and references—its status as fictive, most often (if not always) through a contrasting reference to something in the real-world.

When characters make a metareferential utterance, they remind the audience that the audience is in a real-world context that is separate from the fictive context of the character. This causes the audience to consider the real-world context when they would otherwise engage with the fictive context. This mixing of contexts has the effect of bringing the fictive context into the real world and instantiating it there. Often, especially in film, metareference is used to comedic effect.

Creators of fiction may employ metareference to give an entertaining nod to the audience. Deadpool's mention of McAvoy is a comedic wink to those audience members that have seen other Marvel films and know that he portrays the Professor.

The full line in (4) is "McAvoy or Stewart? These timelines are so confusing." There is another instance of metareference in regard to Stewart (another actor who has played the Professor). More conspicuous is the deictic reference to the X-Men timelines. In any given possible real world, there is only one timeline, so the explicit mentioning of multiple timelines indicates that Deadpool knows he is in a fictive world.

Alternatively, the technique may be used to signal that the fictive context should be mixed with the real-world context, so as to make the fictional world seem more "real" or genuine. In the song "Lane Boy" by Twenty One Pilots, the singer states, "There's a few songs on this record that feel common." By saying this line, the singer reveals that he is aware that he is singing a song, and in doing so brings the song (and the album) closer to the real world inhabited by his audience.

Returning to the discrepancy between (5) and (6), we propose that the audience accommodates some background information more easily than other information [2, 4]. Dorothy lives on a farm, so accommodating that she lives in Kansas is easier than accommodating for Feynman's existence in a fictive world that does not seem to provide any evidence for his existence.

Conclusion

Metareference is a phenomenon of self-awareness in fictional characters. We have provided a list of criteria for determining whether an utterance is metareferential: the utterance does not have an interpretation under the model of the fictive world; the utterance does have an interpretation when the model of the fictive world is mixed with the model of the real world; and the interpretation involves some reference to the fictive material as it stands in the real world. We have shown how this analysis explains the existence or non-existence of metareference in a series of examples.

References

1. Elizabeth Coppock. Semantics boot camp, 2014.

2. Irene Heim. On the projection problem for presuppositions. *Formal semantics–the essential readings*, pages 249–260, 1983.

3. Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142, 1993.

4. Yan Huang. *Pragmatics*. Oxford University Press, 2014.

5. Robert Stalnaker. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457, 1973.

6. Werner Wolf, Walter Bernhart, and Katharina Bantleon. *Metareference across media: theory and case studies: dedicated to Walter Bernhart on the occasion of his retirement*. Rodopi, 2009.