Game theory in semantics and pragmatics

Gerhard Jäger (2013)

Linguistic utterances consist of words and their arrangement within a communicative context where speaker and hearer can reason about each other's knowledge state.

Previous model: epistemic logic to model inferences about mutual belief states

Alternative: Game theory, which focuses on decisions and preferences of agents instead of explicitly modeling internal states and reasoning processes

Nash equilibrium: a solution concept of a non-cooperative game in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only their strategy

Iterated Best Response Model (IBR): derive mapping from *what is said* to *what is meant*

Signaling game model of communication:

1. Sender is assigned private information, its *type*
2. Sender transmits a *signal* to receiver. Choice of signal may depend on type
3. Receiver chooses an *action*, possibly dependent on the observed signal

Players' preferences over histories are modeled as numerical *utilities.*
Rationality dictates that players make choices in a way that maximizes their *expected utilities*, given their epistemic state.

Components of the signaling game:

S	a finite set of possible *signals*
T	a finite set of *types* (information states) the sender might be in
p*	a prior probability distribution over T
truth relation	a truth relation between T and S
A	a set of *actions* that the receiver may take
u_s and u_r	utility functions (sender and receiver) mapping TxSxA to numbers

Do not assume that the receiver considers a certain type more likely than any other type, so if two possibilities *a* and *b* are indistinguishable except for their names, they have the same subjective probability.

Strong interpretation game

$\mathcal{T} = \{t \subseteq \mathcal{S} \mid t \neq \emptyset \wedge t \cup ct \text{ is consistent} \wedge \forall s' \in \mathcal{S}.s' \text{ is consistent with } t \cup ct \Rightarrow s' \in t\},$	Finite set of types that can be expressed as signals consistent with context
$p^*(t) = 1/\lvert\mathcal{T}\rvert,$	Prior distribution over types
$\mathcal{S} = ALT(s),$	Set of signals consistent with context
$t \models s' \Leftrightarrow s' \in t,$	A signal is true in a type iff that signal is an element of t
$\mathcal{A} = \mathcal{T},$	The actions of the receiver are to determine type
$u_s(t, s', a) = -c_s(t, s') + \begin{cases} 1 & \text{if } t = a, \\ 0 & \text{else,} \end{cases}$ $u_r(t, s', a) = -c_r(s', a) + \begin{cases} 1 & \text{if } t = a, \\ 0 & \text{else.} \end{cases}$	Utility functions of sender and receiver

Set of strategies for **sender** (sigma):

$$\Sigma = \mathcal{T} \mapsto \Delta(\mathcal{S}),$$

σ(s|t) is the probability that the sender uses signal s given that she is in type t.

Set of strategies for **receiver** (rho):

$$P = \mathcal{S} \mapsto \Delta(\mathcal{A}),$$

ρ(a|s) is the probability that, upon observing signal s, the receiver will perform action a.

Utility function for **sender**:

$$EU_s(s|t;\rho) = \sum_{a \in \mathcal{A}} \rho(a|s) u_s(t, s, a).$$

which simplifies to:

$$EU_s(s|t;\rho) = \rho(t|s) - c_s(t, s).$$

We calculate the **receiver**'s utility function using Bayes' rule:

$$\sigma(t|s) = \frac{\sigma(s|t)p^*(t)}{\sum_{t' \in \mathcal{T}} \sigma(s|t')p^*(t')}.$$

If p^* is a uniform distribution, this simplifies to

$$\sigma(t|s) = \frac{\sigma(s|t)}{\sum_{t' \in \mathcal{T}} \sigma(s|t')}.$$

The best response of a player to a model of the other player is the behavior strategy that assigns equal probabilities to all best responses and 0 to all sub-optimal responses.

$$BR_s(\rho) = \sigma \quad \text{iff} \quad \sigma(s|t) = \begin{cases} 1/|br_s(t,\rho)| & \text{if } s \in br_s(t,\rho), \\ 0 & \text{else,} \end{cases}$$

$$BR_r(\sigma) = \rho \quad \text{iff} \quad \rho(a|s) = \begin{cases} 1/|br_r(s,\sigma)| & \text{if } a \in br_s(s,\sigma), \\ 0 & \text{else.} \end{cases}$$

For a **receiver** who assumes an honest sender,

$$\sigma_0(s|t) = \begin{cases} 1/|t| & \text{if } s \in t, \\ 0 & \text{else.} \end{cases}$$

	NO	SOME	ALL
$w_{\neg\exists}$	1	0	0
$w_{\exists\neg\forall}$	0	1	0
$w_{\forall}$	0	1/2	1/2

Table 1: Honest sender σ_0

Calculate **receiver**'s utility function using Bayes' rule:

	$w_{\neg\exists}$	$w_{\exists\neg\forall}$	$w_{\forall}$
NO	1	0	0
SOME	0	2/3	1/3
ALL	0	0	1

Table 2: $EU_r(\cdot|\cdot;\sigma_0)$

Then create a best response for each received signal:

	$w_{\neg\exists}$	$w_{\exists\neg\forall}$	$w_{\forall}$
NO	1	0	0
SOME	0	1	0
ALL	0	0	1

Table 3: $\rho_0 = BR_r(\sigma_0)$

Nash equilibrium: If I believe that my opponent has a correct model of my behavior and is rational, I will ascribe to them a best response to my behavior and they ascribe to me a best response to my behavior.

People are not always good at finding these equilibria, but they improve with iterative games.

Iterated Best Response (IBR) sequence:

$$\begin{aligned} \sigma_0(s|t) &= \begin{cases} 1/|t| & \text{if } s \in t, \\ 0 & \text{else,} \end{cases} \\ \rho_n &= BR_r(\sigma_n), \\ \sigma_{n+1} &= BR_s(\rho_n). \end{aligned}$$

Examples

(5) (Who of Ann and Bert came to the party?) Ann came to the party.

Construct a table of σ_0 for an honest sender:

	ANN	BERT	NEITHER	BOTH
w_a	1	0	0	0
w_b	0	1	0	0
$w_\emptyset$	0	0	1	0
w_{ab}	1/3	1/3	0	1/3

Table 5: σ_0

Calculate **receiver**'s utility function:

Then calculate best response for each received signal:

ρ_0	w_a	w_b	$w_\emptyset$	w_{ab}
ANN	1	0	0	0
BERT	0	1	0	0
NEITHER	0	0	1	0
BOTH	0	0	0	1

σ_1	ANN	BERT	NEITHER	BOTH
w_a	1	0	0	0
w_b	0	1	0	0
$w_\emptyset$	0	0	1	0
w_{ab}	0	0	0	1