Computational Pragmatics

Autumn 2015

Raquel Fernández
Institute for Logic, Language & Computation
University of Amsterdam

Outline

Today:

- Part 1: Speech act theory and dialogue acts
  - Homework #2: dialogue acts in the Switchboard corpus
- Part 2: Methodological issue: inter-annotator agreement

Friday:

- Discussion of a recent paper on dialogue act recognition
- Introduction to grounding (negotiating understanding)

Some key units of analysis

(we have already seen)

- Turns: stretches of speech by one speaker bounded by that speaker's silence – that is, bounded either by a pause in the dialogue or by speech by someone else.
- Utterances: units of speech delimited by prosodic boundaries (such as boundary tones or pauses) that form *intentional units* – that is, that can be analysed as an action performed with the intention of achieving something.
- Dialogue acts: intuitively, conversations are made up of sequences of actions such as *questioning, acknowledging*,. . . a notion rooted in *speech act theory*.

Speech Act Theory

Initiated by Austin (*Who to do things with words*) and developed by Searle in the 60s-70s within philosophy of language.

Speech act theory grows out of the following observations:

- Typically, the meaning of a sentence is taken to be its truth value.
- There are utterances for which it doesn't makes sense to say whether they are true or false, e.g., (2)-(5):

> (1) The director bought a new car this year.
> (2) I apologize for being late.
> (3) I promise to come to your talk tomorrow afternoon.
> (4) Put the car in the garage, please.
> (5) Is she a vegetarian?

- These (and generally all) utterances serve to perform actions.
- This is an aspect of meaning that cannot be captured in terms of truth-conditional semantics ($\rightsquigarrow$ felicity conditions).

Types of Acts

What are exactly the actions that are preformed by utterances?
Austin identifies three types of acts that are performed simultaneously:

- locutionary act: basic act of speaking, of uttering a linguistic expression with a particular phonetics/phonology, morphology, syntax, and semantics.
- illocutionary act: the kind of action the speaker intends to accomplish, e.g. *blaming, asking, thanking, joking,...*
  - these functions are commonly referred to as the illocutionary force of an utterance $\rightsquigarrow$ its speech act.
- perlocutionary act: the act(s) that derive from the locution and illocution of an utterance (effects produced on the audience); not always intended and are not under the speaker's control.

John Austin (1962), *How to do things with words*, Oxford: Clarendon Press.

Relations between Acts

Locutionary vs. illocutionary acts:

- The same locutionary act can have different illocutionary forces in different contexts:

> The gun is loaded $\rightsquigarrow$ *threatening? warning? explaining?*

- Conversely, the same illocutionary act can be realised by different locutionary acts:

> Three different ways of carrying out the speech act of requesting:
>
> (6) A day return ticket to Utrecht, please.
>
> (7) Can I have a day return ticket to Utrecht, please?
>
> (8) I'd like a day return ticket to Utrecht.

Key problem: illocutionary acts are a very useful level of abstraction, but how do we map from utterances to speech acts?

Types of Illocutionary Acts

Searle distinguished between five basic types of speech acts:

- Representatives: the speaker is committed to the truth of the expressed proposition (assert, inform)
- Directives: the speaker intends to ellicit a particular action from the hearer (request, order, advice)
- Commissives: the speaker is committed to some future action (promise, oaths, vows)
- Expressives: the speaker expresses an attitude or emotion towards the proposition (congratulations, excuses, thanks)
- Declarations: the speaker changes the reality in accord with the proposition of the declaration (provided certain conventions hold), e.g. baptisms, pronouncing someone guilty.

John Searle (1975), *The Classification of Illocutionary Acts*, Language in Society.

Felicity Conditions

Speech acts are characterised in terms of felicity conditions (rather than truth conditions): conditions under which utterances can be used to properly perform actions (specifications of appropriate use).

Searle identifies four types of felicity conditions (Speaker, Hearer):

Conditions	REQUESTING	PROMISING
propositional content	S intends future act A by H	S intends future act A by S
preparatory	a) S believes H can do A b) It isn't obvious that H would do A without being asked	a) S believes H wants S doing A b) It isn't obvious that S would do A in the normal course of events
sincerity	S wants H to do A	S intends to do A
essential	The utterance counts as an attempt to get H to do A	The utterance counts as an undertaking to do A

Dimensions on which a speech act can go wrong.

Beyond Speech Acts

Speech act theory was developed by philosophers of language (Austin 1962, Searle 1975) $\rightsquigarrow$ their methodology forgoes looking at actual dialogues.

Empirical traditions that have also shaped current dialogue research:

- Conversation Analysis (sociology): Sachs, Schegloff, Jefferson
- Joint Action models (cognitive psychology): Clark, Brennan, ...

Speech act theory focusses on the intentions of the speaker. But a dialogue is not simply a sequence of actions each performed by individual speakers.

- Dialogue is a joint action that requires coordination amongst participants (like playing a duet, dancing a waltz)
  - many actions in dialogue serve to manage the interaction itself
  - they are overlooked by speech act theory
- There are regular patterns of actions that co-occur together

Adjecency Pairs

Certain patterns of dialogue acts are recurrent across conversations

question – answer
proposal – acceptance / rejection / counterproposal
greeting – greeting

Adjacency pairs (term from Conversation Analysis)

- pairs of dialogue act types uttered by different speakers that frequently co-occur in a particular order
- the key idea is not strict adjacency but *expectation*.
  - given the first part of a pair, the second part is immediately relevant and expected (notions of *preferred* and *dispreferred* second parts)
  - intervening turns perceived as an *insertion sequence* or *sub-dialogue*

```
Waitress:  What'll ya have girls?
Customer:  What's the soup of the day?
Waitress:  Clam chowder.
Customer:  I'll have a bowl of clam chowder and a salad.
```

Schegloff (1972), Sequencing in conversational openings, in *Directions in Sociolinguistics*.
Schegloff & Sacks (1973), Opening up closings, *Semiotica*, 7(4):289–327.

The Joint Action Model

Also called collaborative model or grounding model.
[$\hookrightarrow$ *more on grounding this Friday*]

- Clark & Schaefer (1989) put forward a model of dialogue interaction that sees conversation as a joint process, requiring actions by speakers and addressees.
- Speakers and addressees have mutual responsibility for ensuring the success of the communication (need to provide feedback).
- An utterance may have multiple functions at different levels (e.g., asking and giving negative feedback about the communication process)

Clark & Schaefer (1989) Contributing to discourse. *Cognitive Science*, 13:259–294.

Clark (1996) *Using Language*. Cambridge University Press.

From Speech Acts to Dialogue Acts

The concept of dialogue act (DA) extends the notion of speech act to incorporate ideas from conversation analysis and joint action models of dialogue.

It is the term favoured within computational linguistics to refer to the function or the role of an utterance within a dialogue.

- Taxonomies of DAs aim to cover a broader range of utterance functions than traditional speech act types
  - importantly, they include grounding-related DAs (meta-communicative).
- They aim to be effective as tagsets for annotating dialogue corpora.

Dialogue Act Taxonomies: DAMSL

One of the most influential DA taxonomies is the DAMSL schema (Dialogue Act Markup in Several Layers) by Core & Allen (1997).

- Communicative Status
- Information Level
- Forward-looking Function
- Backward-looking Function

Explore the annotation manual:

http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/RevisedManual.html

Utterances can perform several functions at once: possibly one tag per layer.

The taxonomy is meant to be general but not totally domain independent $\rightsquigarrow$ it has been adapted to several types of dialogue.

DA Taxonomies: SWBD DAMSL

The SWBD DAMSL schema is a version of DAMSL created to annotate the Switchboard corpus. Here are the 18 most frequent DA in the corpus:

Tag	Example	Count	%
Statement	*Me, I'm in the legal department.*	72,824	36%
Continuer	*Uh-huh.*	37,096	19%
Opinion	*I think it's great*	25,197	13%
Agree/Accept	*That's exactly it.*	10,820	5%
Abandoned/Turn-Exit	*So, -/*	10,569	5%
Appreciation	*I can imagine.*	4,633	2%
Yes-No-Question	*Do you have to have any special training*	4,624	2%
Non-verbal	*<Laughter>,<Throat_clearing>*	3,548	2%
Yes answers	*Yes.*	2,934	1%
Conventional-closing	*Well, it's been nice talking to you.*	2,486	1%
Uninterpretable	*But, uh, yeah*	2,158	1%
Wh-Question	*Well, how old are you?*	1,911	1%
No answers	*No.*	1,340	1%
Response Ack	*Oh, okay.*	1,277	1%
Hedge	*I don't know if I'm making any sense*	1,182	1%
Declarative Question	*So you can afford to get a house?*	1,174	1%
Other	*Well give me a break, you know.*	1,074	1%
Backchannel-Question	*Is that right?*	1,019	1%

The average conversation consists of 144 turns, 271 utterances, and took 28 min. to annotate. The inter-annotator agreement was 84% (κ=.80).

http://www.stanford.edu/~jurafsky/manual.august1.html

Daniel Jurafsky (2004) Pragmatics and Computational Linguistics. *Handbook of Pragmatics*. Oxford: Blackwell.

Interim Summary

- Speech act theory: truth-conditional content falls short of characterising the role utterance play in conversation. Utterances are actions, with certain *felicity conditions*.
- Conversation analysis / joint action models: we should actually look beyond individual speech acts and embrace the fact that conversations involve multiple participants performing *joint actions* (adjacency pairs, contributions: presentation/response)
- The notion *dialogue act* extends the notion of speech act to incorporate ideas from CA and joint action models.
- DA taxonomies provide inventories of dialogue act types that aim to be suitable for dialogue corpora annotation.

Homework #2

- Investigate two different dialogue act types in the Switchboard Corpus, quantitatively and qualitatively.
- Submission deadline: Friday 18 Sept, 13h.

- There are readily available Python modules for processing the Switchboard Corpus (NLTK and modules by Chris Potts – see homework sheet).
- You are welcome to contact Julian if you have trouble getting started.

Methodology: Inter-annotator agreement

Linguistic Annotation

Important for supervised learning methods and theory validation.

Can we rely on the judgements of one single individual?

> **From Carletta (1996):**
>
> "At one time, it was considered sufficient when working with such judgments to show examples based on the authors' interpretation. Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand, and reliably make, the judgments underlying the research. This is a reasonable requirement, because if researchers cannot even show that people can agree about the judgments on which their research is based, then there is no chance of replicating the research results."

Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249–254.

- an annotation is considered reliable if several annotators agree sufficiently – they consistently make the same decisions.

Inter-annotator Agreement

- Some terminology and notation:
  - set of *items* $\{i \mid i \in I\}$, with cardinality $\mathbf{i}$.
  - set of *categories* $\{k \mid k \in K\}$, with cardinality $\mathbf{k}$.
  - set of *coders* $\{c \mid c \in C\}$, with cardinality $\mathbf{c}$.

Observed Agreement

The simplest measure of agreement is observed agreement A_o:

- the percentage of judgements on which the coders agree, that is the number of items on which coders agree divided by total number of items.

Binary classification task (true / false): rhetorical question

items	**coder A**	**coder B**	**agr**
I mean, why not.	true	true	✓
How do you think we're going to pay for it?	false	true	×
Isn't that sad?	true	false	×
Did you use to live around here?	false	false	✓
Where's that?	false	false	✓
You ever go by Lucky Computer there?	false	false	✓

- $A_o = 4/6 = 66.6\%$

Contingency table:

	coder B		
coder A	true	false	
true	*1*	1	2
false	1	*3*	4
	2	4	6

Contingency table with proportions:
(each cell divided by total # of items **i**)

	coder B		
coder A	true	false	
true	*.166*	.166	.333
false	.166	*.5*	.666
	.333	.666	1

- $A_o = .166 + .5 = .666 = 66.6\%$

Observed vs. Chance Agreement

Problem: using observed agreement to measure reliability does not take into account agreement that is due to chance.

- In the above example, if annotators make random choices the expected agreement due to chance is 50%:
  - both coders randomly choose `true` $(.5 \times .5 = .25)$
  - both coders randomly choose `false` $(.5 \times .5 = .25)$
  - expected agreement by chance: $.25 + .25 = 50\%$
- An observed agreement of 66.6% is only mildly better than 50%

Factors that can lead to higher chance agreement

- Number of categories: fewer categories will result in higher agreement by chance.
  $\mathbf{k} = 2 \rightarrow 50\%$ $\mathbf{k} = 3 \rightarrow 33\%$ $\mathbf{k} = 4 \rightarrow 25\%$ ...

- Distribution of items among categories: if some categories are very frequent, observed agreement will be higher by chance.
  - both coders randomly choose `true` $(.95 \times .95 = 90.25\%)$
  - both coders randomly choose `false` $(.05 \times .05 = 0.25\%)$
  - expected agreement by chance $90.25 + 0.25 = 90.50\%$

  $\Rightarrow$ Observed agreement of 90% may be less than chance agreement.

In sum, observed agreement does not take chance agreement into account and hence is not a good measure of reliability.

Measuring Reliability

$\Rightarrow$ Reliability measures must be corrected for chance agreement.

- Let A_o be observed agreement, and A_e expected agreement by chance.
- $1 - A_e$: how much agreement beyond chance is attainable.
- $A_o - A_e$: how much agreement beyond chance was found.
- General form of chance-corrected agreement measure of reliability:

$$R = \frac{A_o - A_e}{1 - A_e}$$

The ratio between $A_o - A_e$ and $1 - A_e$ tells us which proportion of the possible agreement beyond chance was actually achieved.

- Some general properties of R:

perfect agreement	**chance agreement**	**perfect disagreement**
$R = 1 = \frac{A_o - A_e}{1 - A_e}$	$R = 0 = \frac{0}{1 - A_e}$	$R = \frac{0 - A_e}{1 - A_e}$

Measuring Reliability: *kappa*

Several agreement measures have been proposed in the literature

Arstein & Poesio (2008) Survey Article: Inter-Coder Agreement for Computational Linguistics, *Computational Linguistics*, 34(4):555–596.

- The general form of R is the same for several measures $R = \frac{A_o - A_e}{1 - A_e}$
- They all compute A_o in the same way:
  - proportion of agreements over total number of items
- They differ on the precise definition of A_e.

We'll focus on the kappa (κ) coefficient (Cohen 1960; see also Carletta 1996)

- κ calculates A_e considering *individual* category distributions:
  - they can be read off from the marginals of contingency tables:

	coder B		
coder A	true	false	
true	1	1	2
false	1	3	4
	2	4	6

	coder B		
coder A	true	false	
true	.166	.166	*.333*
false	.166	.5	*.666*
	.333	.666	1

category distribution for coder A: $P(\texttt{true}|c_A) = .333$; $P(\texttt{false}|c_a) = .666$
category distribution for coder B: $P(\texttt{true}|c_B) = .333$; $P(\texttt{false}|c_B) = .666$

Chance Agreement for *kappa*

A_e: how often are annotators expected to agree if they make random choices according to their individual category distributions?

- we assume that the decisions of the coders are independent: need to multiply the marginals
- Chance of c_A and c_B agreeing on category k: $P(k|c_A) \cdot P(k|c_B)$
- A_e is then the chance of the coders agreeing on any k:

$$A_e = \sum_{k \in K} P(k|c_A) \cdot P(k|c_B|)$$

	coder B		
coder A	true	false	
true	1	1	2
false	1	3	4
	2	4	6

	coder B		
coder A	true	false	
true	.166	.166	*.333*
false	.166	.5	*.666*
	.333	.666	1

- $A_e = (.333 \cdot .333) + (.666 \cdot .666) = .111 + .444 = 55.5\%$

An Example

items	coder A	coder B	agr
I mean, why not.	true	true	✓
How do you think we're going to pay for it?	false	true	×
Isn't that sad?	true	false	×
Did you use to live around here?	false	false	✓
Where's that?	false	false	✓
You ever go by Lucky Computer there?	false	false	✓

	coder B		
coder A	true	false	
true	1	1	2
false	1	3	4
	2	4	6

	coder B		
coder A	true	false	
true	.166	.166	*.333*
false	.166	.5	*.666*
	.333	.666	1

- $A_o = .166 + .5 = .666 = 66.6\%$
- $A_e = (.333 \cdot .333) + (.666 \cdot .666) = .111 + .444 = 55.5\%$

$$\kappa = \frac{66.6 - 55.5}{100 - 55.5} = \frac{11.1}{44.5} = \mathbf{24.9\%}$$

kappa for more than two coders

kappa can be generalised to multiple coders.

- We need to compute pairwise observed agreement:
  - the amount of agreement for each item i is the proportion of agreeing pairwise judgements out of the total number of pairwise judgments for i.
  - A_o is the mean of the amount of agreement for all items $i \in I$
- We need to compute pairwise expected agreement:
  - recall that k uses the individual category distributions $P(k|c) = \mathbf{n}_{kc}/\mathbf{i}$
  - the chance of two coders agreeing on k is $P(k|c_A) \cdot P(k|c_B)$
  - the chance of two arbitrary coders c_n and c_m agreeing on category k is the mean of $P(k|c_n) \cdot P(k|c_m)$ over all pairs of coders.
  - A_e is the sum of this join probability over all $k \in K$.
  - (this is equivalent to the mean of A_e for all pairs of coders)

Scales for the Interpretation of Kappa

- Landis and Koch (1977)

```
0.0 – 0.2 :    slight
0.2 – 0.4 :    fair
0.4 – 0.6 :    moderate
0.6 – 0.8:   substantial
0.8 – 1.0 :  perfect
```

- Krippendorff (1980)

```
0.0 – 0.67 :    discard
0.67 – 0.8 :  tentative
0.8 – 1.0:  good
```

- Green (1997)

```
0.0 – 0.4 :    low
0.4 – 0.75 :  fair / good
0.75 – 1.0:  high
```

- There are many other suggestions as well. . .

Weighted Disagreements

- The classic version of κ considers all types of disagreements equally.
- However, we may want to treat some disagreements as more important than others – some categories may be more similar than others.
- We can use weighted coefficients: Krippendorff's α and *weighted kappa* κ_w.
  - The formula for κ_w derives agreement from disagreement:

$$\kappa_w = 1 - \frac{D_o}{D_e}$$

  - We'll see how to derive D_o and D_e from the confusion matrices; for details of the formulas see Arstein & Poesio (2008).

Weighted Disagreements – An Example

Consider this confusion matrix from Arstein & Poesio (2008):

	coder B			
coder A	Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

We can calculate unweighted κ as described before:

- A_o : the sum of the cells in the diagonal
  $A_o = .46 + .32 + .10 = .88$
- A_e : the sum of the marginals for each category (multiplied)
  $A_e = .46 \times .52 + .44 \times .32 + .10 \times .16 = .396$
- $\kappa = (A_o - A_e)/(1 - A_e)$
  $\kappa = (.88 - .396)/(1 - .396) = .8013$

Weighted Disagreements – An Example

Suppose we weight the distances between the categories as shown in the RHS table: identical categories have 0 disagreement, while 1 denotes maximal disagreement.

	coder B			
coder A	Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

	coder B		
coder A	Stat	IReq	Chck
Stat	0	1	0.5
IReq	1	0	0.5
Chck	0.5	0.5	0

To calculate κ_w, we can derive D_o and D_e as follows:

- D_o : the sum of all cells multiplying each cell by each weight (and dividing by total of items if not working with proportions).
- D_e : the sum of $D_e^{k_i k_j}$ for each category pair k_i, k_j, where
  - $D_e^{k_i k_j}$: the product of the marginals for k_i and k_j divided by the total of items (or the square of the total of items if not working with proportions), multiplying each cell by each weight.

Weighted Disagreements – An Example

coder A	coder B: Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

coder A	coder B: Stat	IReq	Chck
Stat	0	1	0.5
IReq	1	0	0.5
Chck	0.5	0.5	0

- D_o : the sum of all cells multiplying each cell by each weight (and dividing by total of items if not working with proportions).

$$\mathrm{D_o} = \frac{46\times 0 + 6\times 1 + 32\times 0 + 6\times 0.5 + 10\times 0}{100} = \frac{6+3}{100} = 0.09$$

- D_e : the sum of $D_e^{k_i k_j}$ for each category pair k_i, k_j, where
  - $D_e^{k_i k_j}$: the product of the marginals for k_i and k_j divided by the total of items (or the square of the total of items if not working with proportions), multiplying each cell by each weight.

$$\begin{aligned} &\tfrac{46\times 52}{100\times 100}\times 0 + \tfrac{44\times 52}{100\times 100}\times 1 + \tfrac{10\times 52}{100\times 100}\times \tfrac{1}{2} \\ &+ \tfrac{46\times 32}{100\times 100}\times 1 + \tfrac{44\times 32}{100\times 100}\times 0 + \tfrac{10\times 32}{100\times 100}\times \tfrac{1}{2} \\ &+ \tfrac{46\times 16}{100\times 100}\times \tfrac{1}{2} + \tfrac{44\times 16}{100\times 100}\times \tfrac{1}{2} + \tfrac{10\times 16}{100\times 100}\times 0 \end{aligned} \qquad 0.49$$

Weighted Disagreements – An Example

	coder B			
coder A	Stat	IReq	Chck	
Stat	46	6	0	52
IReq	0	32	0	32
Chck	0	6	10	16
	46	44	10	100

	coder B		
coder A	Stat	IReq	Chck
Stat	0	1	0.5
IReq	1	0	0.5
Chck	0.5	0.5	0

$$\kappa_w = 1 - \frac{D_o}{D_e}$$

$\kappa_w = 1 - (.09/.49) = .8163$

$\kappa = (.88 - .396)/(1 - .396) = .8013$

Different types of non-reliability

- Misinterpretation of annotation guidelines: may not result in disagreement $\rightarrow$ may not be detected
- Random slips: lead to chance agreement between annotators
- Different intuitions: lead to systematic disagreements

Gold-standard annotations

An annotated linguistic corpus typically is released with 1 annotation, considered the gold standard.

- often only a small part of the annotation is tested for reliability
- experts make tricky decisions
- annotators discuss and reach a consensus

The construction of annotated linguistic resources is rapidly and radically changing with the use of crowdsourcing platforms, like Amazon's Mechanical Turk.

- motivation: cheap, quick, more data
- challenges: how do we make sure the annotation is reliable? how to derive a gold standard?

Our website on Collective Annotation `http://www.illc.uva.nl/Resources/CollectiveAnnotation/`