Mechanisms of Meaning

Autumn 2010

Raquel Fernández

Institute for Logic, Language & Computation
University of Amsterdam

Plan for Today

- *Part 1*: Assessing the reliability of linguistic annotations with inter-annotator agreement
  - discussion of the semantic annotation exercise
- *Part 2*: Psychological theories of concepts and word meaning
  - presentation and discussion of chapter 2 of Murphy (2002): *Typicality and the Classical View of Categories*

- *Next week*: Presentation and discussion of Murphy's
  - chapter 3: *Theories* (by Marta Sznajder)
  - chapter 11: *Word Meaning* (by Adam Pantel)

Semantic Judgements

Theories of linguistic phenomena are typically based on speakers' *judgements* (regarding e.g. acceptability, semantic relations, etc.).

As an example, consider a theory that proposes to predict different dative structures from different senses of *'give'*.

- Hypothesis: different conceptualisations of the giving event are associated with different structures [refuted by Bresnan et al. 2007]

causing a change of state (possession)	⇒	V NP NP *Susan gave the children toys*
causing a change of place (movement to a goal)	⇒	V NP [to NP] *Susan gave toys to the children*

- Some evidence for this hypothesis comes from give idioms:

That movie gave me the creeps	/ *gave the creeps to me
That lighting gives me a headache	/ *gives a headache to me

Bresnan et al. (2007) Predicting the Dative Alternation, *Cognitive Foundation of Interpretation*, Royal Netherlands Academy of Arts and Sciences.

Semantic Judgements

What do we need to confirm this hypothesis? At least, the following:

- data: a set of *'give'* sentences with different dative structures;
- judgements indicating the type of giving event in each sentence.

This raises several issues, among others:

- how much data? what kind of data - constructed examples?
- whose judgements? the investigator's? those of native speakers - how many? what if judgements differ among speakers?

How to overcome the difficulties associated with semantic judgements?

- Possibility 1: forget about judgements and work with raw data
- Possibility 2: take judgements from several speakers, measure their agreement, and aggregate them in some meaningful way.

Annotations and their Reliability

When data and judgements are stored in a computer-readable format, judgements are typically called *annotations*.

- What are linguistic annotations useful for?
  - they allow us to check automatically whether hypotheses relying on particular annotations hold or not.
  - they help us to develop and test algorithms that use the information from the annotations to perform practical tasks.
- Researchers who wish to use manual annotations are interested in determining their *validity*.
- However, since annotations correspond to speakers' judgements, there isn't an objective way of establishing validity...
- Instead, measure the *reliability* of an annotation:
  - annotations are reliable if annotators agree *sufficiently for relevant purposes* – they consistently make the same decisions.
  - high reliability is a prerequisite for validity.

Annotations and their Reliability

How can the reliability of an annotation be determined?

- several coders annotate the same data with the same guidelines
- calculate *inter-annotator agreement*

Main references for this topic:

* Arstein an Poesio (2008) Survey Article: Inter-Coder Agreement for Computational Linguistics, *Computational Linguistics*, 34(4):555–596.
* Slides by Gemma Boleda and Stefan Evert part of the ESSLLI 2009 course "Computational Lexical Semantics":
  http://clseslli09.files.wordpress.com/2009/07/02_iaa-slides1.pdf

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 $\mathbf{c}$.
- In our semantic annotation exercise:
  - items: 70 sentences containing two highlighted nouns.
  - categories: *true* and *false*
  - coders: you (+ the SemEval annotators)

items	coder A	coder B	agr
Put *tea* in a *heat-resistant jug* and ...	true	true	✓
The *kitchen* holds patient *drinks* and snacks.	true	false	×
Where are the *batteries* kept in a *phone*?	true	false	×
...the *robber* was inside the *office* when ...	false	false	✓
Often the *patient* is kept in the *hospital* ...	false	false	✓
Batteries stored in *contact* with one another...	false	false	✓

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.

items	coder A	coder B	agr
Put *tea* in a *heat-resistant jug* and ...	true	true	✓
The *kitchen* holds patient *drinks* and snacks.	true	false	×
Where are the *batteries* kept in a *phone*?	true	false	×
...the *robber* was inside the *office* when ...	false	false	✓
Often the *patient* is kept in the *hospital* ...	false	false	✓
Batteries stored in *contact* with one another...	false	false	✓

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

Contingency table:

	coder B		
coder A	true	false	
true	1	2	3
false	0	3	3
	1	5	6

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

	coder B		
coder A	true	false	
true	.166	.333	.5
false	0	.5	.5
	.166	.833	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 our task, 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%

Observed vs. Chance Agreement

Factors that vary across studies and need to be taken into account:

- *Number of categories*: fewer categories will result in higher observed 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` $(.0 \times .05 = 0.25\%)$
  - expected agreement by chance $90.25 + 0.25 = 90.50\%$

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

Observed agreement does not take these factors 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 $A_o - A_e$ 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 (see Arstein & Poesio 2008 for details)

- The general form of R is the same for all 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	2	3
false	0	3	3
	1	5	6

	coder B		
coder A	true	false	
true	.166	.333	.5
false	0	.5	.5
	.166	*.833*	1

category distribution for coder A: $P(c_A|\texttt{true}) = .5$; $P(c_a|\texttt{false}) = .5$
category distribution for coder B: $P(c_B|\texttt{true}) = .166$; $P(c_B|\texttt{false}) = .833$

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: $P(c_A|k) \cdot P(c_B|k)$
- A_e is then the chance of the coders agreeing on any k:

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

	coder B		
coder A	true	false	
true	1	2	3
false	0	3	3
	1	5	6

	coder B		
coder A	true	false	
true	.166	.333	.5
false	0	.5	.5
	.166	*.833*	1

- $A_e = (.5 \cdot .166) + (.5 \cdot .833) = .083 + .416 = 49.9\%$

Kappa for our Example

items	coder A	coder B	agr
Put *tea* in a *heat-resistant jug* and ...	true	true	✓
The *kitchen* holds patient *drinks* and snacks.	true	false	×
Where are the *batteries* kept in a *phone*?	true	false	×
...the *robber* was inside the *office* when ...	false	false	✓
Often the *patient* is kept in the *hospital* ...	false	false	✓
Batteries stored in *contact* with one another...	false	false	✓

	coder B		
coder A	true	false	
true	1	2	3
false	0	3	3
	1	5	6

	coder B		
coder A	true	false	
true	**.166**	.333	.5
false	0	**.5**	.5
	.166	.833	1

- $A_o = .166 + .5 = .666 = 66.6\%$
- $A_e = (.5 \cdot .166) + (.5 \cdot .833) = .083 + .416 = 49.9\%$

$$\kappa = \frac{66.6 - 49.9}{1 - 49.9} = \frac{16.7}{50.1} = \mathbf{33.3}\%$$

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. . .

Semantic Annotation Exercise

- Task 4 at SemEval-2007: Classification of Semantic Relations between Nominals
  - the dataset is meant to be a benchmark for evaluating semantic relation classification algorithms
  - potential application is information retrieval, summarisation, machine translation, ...
- We'll compute κ for each annotator and the *gold standard* provided by SemEval-2007.
  - data set independently annotated by two codes, who examined their disagreements and arrived at a consensus.
- *Kappa* for multiple annotators: compute κ for each possible pair of annotators, then report average (and standard deviation).

Semantic Annotation Exercise

	alessandra	
gold	false	true
false	0.443	0.129
true	0.057	0.371

$A_o = .814$; $A_e = .5$

$\kappa = .628$

	andreas	
gold	false	true
false	0.486	0.086
true	0.029	0.4

$A_o = .885$; $A_e = .502$

$\kappa = .77$

	holger	
gold	false	true
false	0.486	0.086
true	0.129	0.3

$A_o = .785$; $A_e = .516$

$\kappa = .556$

	irma	
gold	false	true
false	0.371	0.2
true	0.086	0.343

$A_o = .714$; $A_e = .493$

$\kappa = .435$

	marta	
gold	false	true
false	0.486	0.086
true	0.1	0.329

$A_o = .814$; $A_e = .512$

$\kappa = .619$

	noortje	
gold	false	true
false	0.471	0.1
true	0.1	0.329

$A_o = .8$; $A_e = .510$

$\kappa = .591$

Average $\kappa = .608$

Semantic Annotation Exercise

andreas	alessandra: false	alessandra: true
false	.414	.1
true	.086	4

$A_o = .814$; $A_e = .5$

$\kappa = .628$

holger	alessandra: false	alessandra: true
false	.443	.171
true	.057	.329

$A_o = .771$; $A_e = .5$

$\kappa = .542$

marta	alessandra: false	alessandra: true
false	.429	.157
true	.071	.343

$A_o = .771$; $A_e = .5$

$\kappa = .542$

noortje	alessandra: false	alessandra: true
false	.429	.143
true	.071	.357

$A_o = .785$; $A_e = .5$

$\kappa = .571$

andreas	holger: false	holger: true
false	.457	.057
true	.157	.329

$A_o = .785$; $A_e = .503$

$\kappa = .568$

andreas	irma: false	irma: true
false	.371	.143
true	.086	.4

$A_o = .771$; $A_e = .498$

$\kappa = .543$

Semantic Annotation Exercise

	marta	
andreas	false	true
false	.414	.1
true	.171	.314

$A_o = .728$; $A_e = .502$

$\kappa = .454$

	noortje	
andreas	false	true
false	. 471	.043
true	.1	.386

$A_o = .857$; $A_e = .502$

$\kappa = .713$

	irma	
holger	false	true
false	.414	.2
true	.043	.343

$A_o = .757$; $A_e = .490$

$\kappa = .523$

	marta	
holger	false	true
false	.5	.114
true	.086	.3

$A_o = .8$; $A_e = .519$

$\kappa = .583$

	noortje	
holger	false	true
false	.471	.143
true	.1	.286

$A_o = .757$; $A_e = .516$

$\kappa = .497$

	marta	
irma	false	true
false	.371	.086
true	.214	.329

$A_o = .7$; $A_e = .492$

$\kappa = .408$

Semantic Annotation Exercise

	noortje	
irma	false	true
false	.4	.057
true	.171	.371

$A_o = .771$; $A_e = .493$

$\kappa = .548$

	noortje	
marta	false	true
false	.457	.129
true	.114	.3

$A_o = .757$; $A_e = .512$

$\kappa = .502$

	alessandra	
irma	false	true
false	.329	.129
true	.171	.371

$A_o = .7$; $A_e = .5$

$\kappa = .4$

Average $\kappa = .534$

Different Types of Non-reliability

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

References

- Artstein, Ron and Poesio, Massimo (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
- Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249–254.
- Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Green, Annette M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual SAS Users Group International Conference*, San Diego, CA.
- Krippendorff, Klaus (1980). *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.
- Landis, J. Richard and Koch, Gary G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Psychological Theories of Concepts and Word Meaning