Tilburg University

Evaluating Dialogue Act Tagging

with Naive & Expert annotators

Jeroen Geertzen & Volha Petukhova & Harry Bunt

LREC 2008 / Marrakech / May 28th

Evaluating dialogue act schemes I

- A dialogue act scheme should be reliable in application:

  assignment of the categories does not depend on individual judgement, but on shared understanding of what the categories mean and how they are to be used.

[1] (Cohen, 1960; Carletta, 1996)

Evaluating dialogue act schemes I

- A dialogue act scheme should be reliable in application:

  assignment of the categories does not depend on individual judgement, but on shared understanding of what the categories mean and how they are to be used.

- Reliability is often evaluated using inter-annotator agreement:
  - Observed agreement (p_o);
  - Standard kappa[1] taking expected agreement (p_e) into account:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

[1] (Cohen, 1960; Carletta, 1996)

Evaluating dialogue act schemes II

- But what kind of annotators to use: naive (NC) or expert (EC) coders?
  - Carletta: for subjective codings there are no real experts
  - Krippendorf[2], Carletta: that what counts is how totally naive coders manage based on written instructions.

[2] (Krippendorf, 1980)

Evaluating dialogue act schemes II

- But what kind of annotators to use: naive (NC) or expert (EC) coders?
  - Carletta: for subjective codings there are no real experts
  - Krippendorf[2], Carletta: that what counts is how totally naive coders manage based on written instructions.
- For naive coders, factors such as instruction clarity or annotation platform have more impact
- Using expert coders makes sense with complex tagsets and when aiming for as-accurate-as-possible annotations

[2] (Krippendorf, 1980)

Research question

- Annotation by *both* NC and EC are insightful:
  - NC: insight in clarity of concepts
  - EC: reliability when errors due to conceptual misunderstanding and lack of experience are minimized

Research question

- Annotation by *both* NC and EC are insightful:
    - NC: insight in clarity of concepts
    - EC: reliability when errors due to conceptual misunderstanding and lack of experience are minimized
- How do both annotator groups differ in annotating?
    - $\Longrightarrow$ contrast NC annotations with EC annotations and evaluate on both *inter annotator agreement (IAA)* and *tagging accuracy (TA)*
    - $\Longrightarrow$ qualitative analysis of observed differences

Experiment outline I

- Naive coders:
    - 6 undergraduate students, not linguistically trained
    - 4 hour session explaining data, tagset, and annotation platform
- Expert coders:
    - 2 PhD students, not linguistically trained
    - working with the scheme for more than two years
- Data consisted of task-oriented dialogue in Dutch:

corpus	domain	type	#utt
OVIS	train connections	H-M	193
DIAMOND	operating a fax machine	H-M	131
		H-H	114
DUTCH MAPTASK	map task	H-H	120
			558

Experiment outline II

- Gold standard:
  - established agreement by 3 experts (all authors)
  - few cases with fundamental disagreement / unclarity excluded

Experiment outline II

- Gold standard:
  - established agreement by 3 experts (all authors)
  - few cases with fundamental disagreement / unclarity excluded
- Dialogue act tagset, DIT^{++}:
  - Comprehensive, also containing concepts from other schemes
  - Clearly defined notion of dimension; fine-grained feedback acts
  - In each of the 11 dimensions a specific aspect of communication can be addressed:
    Task, Auto-feedback, Allo-feedback, Own Communication, Partner Communication, Turn, Contact, Time, Dialogue Structuring, Topic, and Social Obligations.
  - For each dimension, at most one act can be assigned.

Results on inter annotator agreement

	naive annotators				expert annotators			
Dimension	p_o	p_e	κ_{tw}	ap-r	p_o	p_e	κ_{tw}	ap-r
task	0.63	0.17	**0.56**	0.81	0.85	0.16	**0.82**	0.78
auto feedback	0.67	0.48	**0.36**	0.53	0.92	0.57	**0.82**	0.64
allo feedback	0.53	0.29	**0.33**	0.02	0.85	0.24	**0.81**	0.38
time	0.87	0.84	**0.20**	0.51	0.98	0.87	**0.88**	0.89
contact	0.80	0.66	**0.41**	0.19	0.75	0.38	**0.60**	0.50
dialogue struct.	0.80	0.30	**0.71**	0.32	0.92	0.38	**0.88**	0.65
social obl.	0.95	0.28	**0.93**	0.72	0.93	0.24	**0.91**	0.86

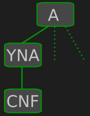
Results on inter annotator agreement

	naive annotators				expert annotators			
Dimension	p_o	p_e	κ_{tw}	ap-r	p_o	p_e	κ_{tw}	ap-r
task	0.63	0.17	**0.56**	0.81	0.85	0.16	**0.82**	0.78
auto feedback	0.67	0.48	**0.36**	0.53	0.92	0.57	**0.82**	0.64
allo feedback	0.53	0.29	**0.33**	0.02	0.85	0.24	**0.81**	0.38
time	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.88**	0.89
contact	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.60**	0.50
dialogue struct.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.88**	0.65
social obl.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.91**	0.86

Taxonomically weighted kappa :

		C1	C2	C3
A	is this correct?			
B	yes	A	YNA	CNF

C1 and C2 show more partial agreement than C1 and C3



Results on inter annotator agreement

	naive annotators				expert annotators			
Dimension	p_o	p_e	κ_{tw}	ap-r	p_o	p_e	κ_{tw}	ap-r
task	0.63	0.17	**0.56**	0.81	0.85	0.16	**0.82**	0.78
auto feedback	0.67	0.48	**0.36**	0.53	0.92	0.57	**0.82**	0.64
allo feedback	0.53	0.29	**0.33**	0.02	0.85	0.24	**0.81**	0.38
time	0.87	0.84	**0.20**	0.51	0.98	0.87	**0.88**	0.89
contact	0.80	0.66	**0.41**	0.19	0.75	0.38	**0.60**	0.50
dialogue struct.	0.80	0.30	**0.71**	0.32	0.92	0.38	**0.88**	0.65
social obl.	0.95	0.28	**0.93**	0.72	0.93	0.24	**0.91**	0.86

Results on tagging accuracy

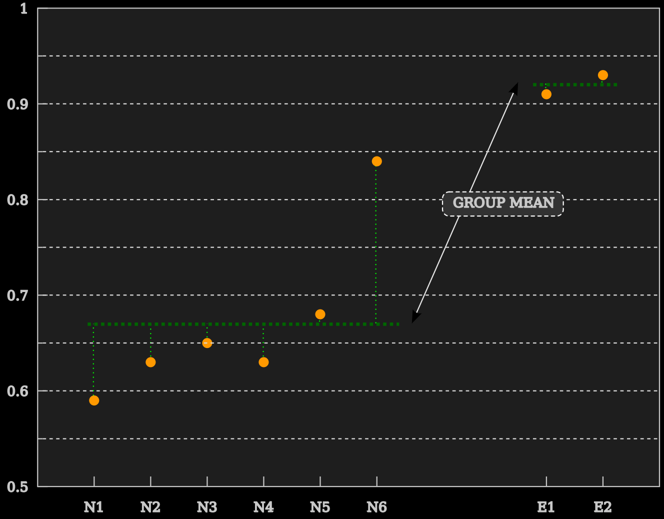
	naive annotators			expert annotators		
Dimension	p_o	p_e	κ_{tw}	p_o	p_e	κ_{tw}
task	0.64	0.16	**0.58**	0.91	0.16	**0.90**
auto feedback	0.74	0.46	**0.52**	0.94	0.48	**0.88**
allo feedback	0.58	0.19	**0.48**	0.95	0.22	**0.94**
time	0.92	0.81	**0.57**	0.99	0.88	**0.94**
contact	1.00	0.60	**1.00**	0.91	0.48	**0.83**
dialogue struct.	0.89	0.36	**0.82**	0.87	0.34	**0.81**
social obl.	0.96	0.26	**0.94**	0.95	0.23	**0.94**

Results on tagging accuracy

	naive annotators			expert annotators		
Dimension	p_o	p_e	κ_{tw}	p_o	p_e	κ_{tw}
task	0.64	0.16	**0.58**	0.91	0.16	**0.90**
auto feedback	0.74	0.46	**0.52**	0.94	0.48	**0.88**
allo feedback	0.58	0.19	**0.48**	0.95	0.22	**0.94**
time	0.92	0.81	**0.57**	0.99	0.88	**0.94**
contact	1.00	0.60	**1.00**	0.91	0.48	**0.83**
dialogue struct.	0.89	0.36	**0.82**	0.87	0.34	**0.81**
social obl.	0.96	0.26	**0.94**	0.95	0.23	**0.94**

- When generalising over all dimensions & calculating a single accuracy score for each group, naive annotators score 0.67 and experts score 0.92

Individual scores of annotators



Observations I

- Sometimes, NC showed less disagreement than EC
- Example for co-occurrence WH-ANSWER - INSTRUCT:

	utterance	expert 1	expert 2
S_1	do you want an overview of the codes?	YN-Q	YN-Q
U_1	yes	YN-A	YN-A
S_2	press function	INSTRUCT	WH-A
S_3	press key 13	INSTRUCT	WH-A
S_4	a list is being printed	INFORM	WH-A

- Where NC followed question-answer adjacency pairs, EC generally disagreed on specificity

Observations II

- In general, and specifically in turn-management, EC recognised multi-functionality more than NC
- Example:

	utterance	naive	expert
A_1	to the left...	TAS:WH-A	TAS:WH-A TUM:KEEP
A_2	**and then** slightly around	TAS:WH-A	TAS:WH-A TUM:KEEP

Conclusions

- Codings by both NC and EC provide complementary insights

Conclusions

- Codings by both NC and EC provide complementary insights
- Calculating TA requires a ground truth, which can be established when concepts are not too subjective

Conclusions

- Codings by both NC and EC provide complementary insights
- Calculating TA requires a ground truth, which can be established when concepts are not too subjective
- NC disagree more (with each other and gold standard) whether or not to annotate in a specific dimension

Conclusions

- Codings by both NC and EC provide complementary insights
- Calculating TA requires a ground truth, which can be established when concepts are not too subjective
- NC disagree more (with each other and gold standard) whether or not to annotate in a specific dimension
- EC show more agreement on when to annotate in a specific dimension, but as a result are also addressing more difficult cases

Conclusions

- Codings by both NC and EC provide complementary insights
- Calculating TA requires a ground truth, which can be established when concepts are not too subjective
- NC disagree more (with each other and gold standard) whether or not to annotate in a specific dimension
- EC show more agreement on when to annotate in a specific dimension, but as a result are also addressing more difficult cases
- Distinguishing agreement on whether or not to annotate in a dimension from agreement on the dialogue act within a dimension is essential

Thanks for your attention !

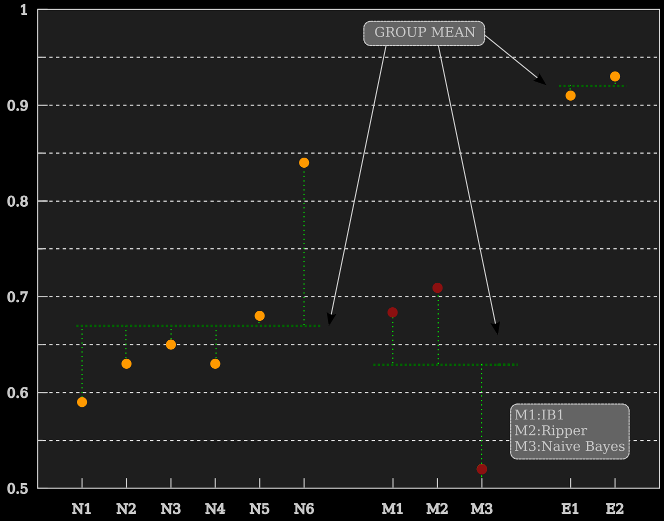
Any questions ?

Announcement:

8th International Conference on Computational Semantics
January 7-9 2009, Tilburg, The Netherlands
Submission deadlines: 1 Oct (long papers) & 27 Oct (short papers)

See: `iwcs.uvt.nl`

Comparing NC and EC with machine learners



M1:IB1
M2:Ripper
M3:Naive Bayes