(Incremental) Dialogue Act Segmentation and Recognition

Volha (Olga) Petukhova

Spoken Language Systems Group
Saarland University

Proseminar Computational Pragmatics
Winter 2019/2020

Outline

- Introduction
- Segmentation
- Classification
- Experimental designs
- Results
- Experiments
- Discussion

Segmentation:
what is the smallest meaningful dialogue unit?

Turn?

turn can be defined as a stretch of communicative behaviour produced by one speaker, bounded by periods of inactivity of that speaker or by activity of another speaker

But:

A1: Well we can chat away for ... um... for five minutes or

so I think at...

B: Mm-hmm

A1: ... at most

Segmentation:
what is the smallest meaningful dialogue unit?

Or

A1: Like you said time to market was a problem and how many components are physically in there in cost

A2: um (0.4)

A3: 0.28 and (0.12) the power is basically a factor of that

A4: 0.55 um (0.47)

A5: and (0.32) the lower components: the power, the logic, the transmitter and the infrared, they affect you in terms of the size of your device

A6: 0.59 um (0.26)

A7: and (0.16) that would have some impact on how y i think more hold rather than the actual use the remote control

Segmentation:

what is the smallest meaningful dialogue unit?

Utterance?

Utterances, on the other hand, are linguistically defined stretches of communicative behaviour that have one or multiple communicative functions

But

*About half ... **about a quar- ... th- ...third** of the way down I have some hills*

*Because twenty five Euros for a remote... **how much is that locally in pounds?** is too much money to buy an extra remote or a replacement remote*

U: *W*hat time is the first train to the airport on Sunday?
S: ***The first train to the airport on Sunday*** is at ...ehm... 6.17.

Segmentation:
what is the smallest meaningful dialogue unit?

Segmentation of spoken dialogue is nontrivial due to phenomena such as:

- filled/unfilled pauses, stalling
- restarts, self-corrections
- phrasal interjections
- interruption & continuation

As a consequence, a meaningful unit can be:

- discontinuous
- spread over multiple turns

Segmentation:

what is the smallest meaningful dialogue unit?

As meaningful units we prefer not to use the notion of utterance, but that of what we call functional segment:

a (possibly discontinuous) stretch of communicative behaviour that has one or multiple communicative functions

Functional segments:

discontinuity

- Set Question

U : what time is the first train to the airport on Sunday?

- Set Answer

S : *at ...ehm... 6.17*

Set Answer STALL

S : *at ...ehm... 6.17*

STALL

Set Answer

Functional segments:
overlapping

- Set Question

U : what time is the first train to the airport on Sunday?

- Set Answer

S : *the first train to the airport on Sunday is at 6.17*

FEEDBACK

Set Answer

no single segmentation exists that indicates the relevant functional segments

Multidimensional segmentation

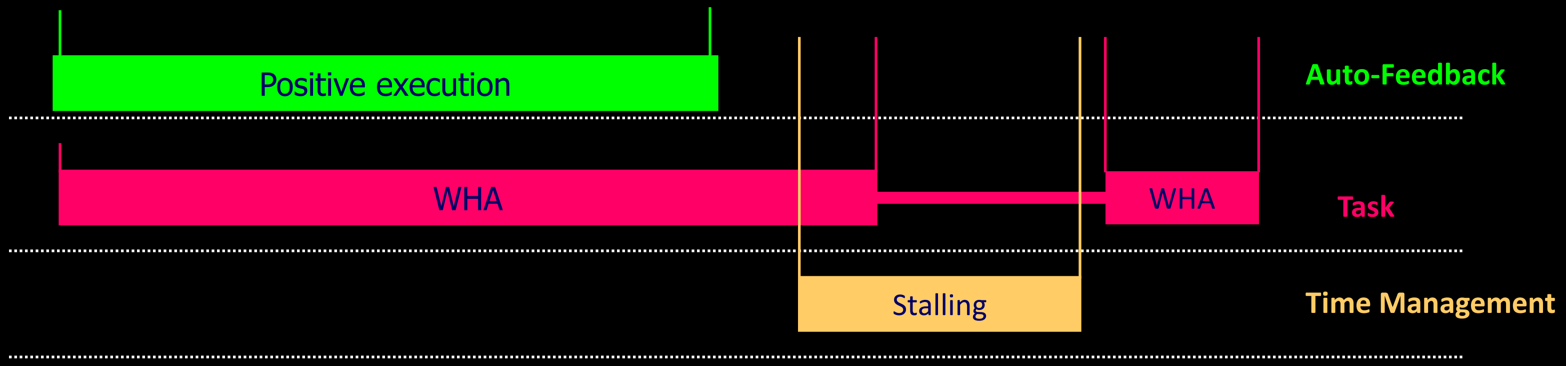
- Solution: use multiple segmentations instead of a single one
- Allows indication of multiple functional segments in an utterance to be identified more accurately
- Compatible with DA taxonomies that address several aspects (‘dimensions’) of dialogue simultaneously (e.g. DAMSL or DIT)

Multidimensional segmentation:

example

U : what time is the first train to the airport on Sunday? (WHQ)

S : the first train to the airport on Sunday is a t ...ehm... 6.17



Automatic segmentation: data and features

- Data

AMI meeting corpus: 3 dialogues with 4 participants: 17,335 words, 504 speaker turns, 1,903 utterances, 3,897 functional segments; average utterance length – 9 words, average segments length is 4.4 words, average turn length of 3.8 utterances and 7.7 segments

- Features:

dialogue history: tags of the 10 (AMI) and 4 previous turns prosody: min/max/avg/stdev of pitch and energy, voicing (fraction of locally unvoiced frames and number of voice breaks), and duration

word occurrence: bag-of-words vector

relations between functional segments

Automatic segmentation: labeling

- Labels for segment boundaries
- BIO labeling: B – begins segment; I – inside segment; O – ends segment; BO – one-token segment

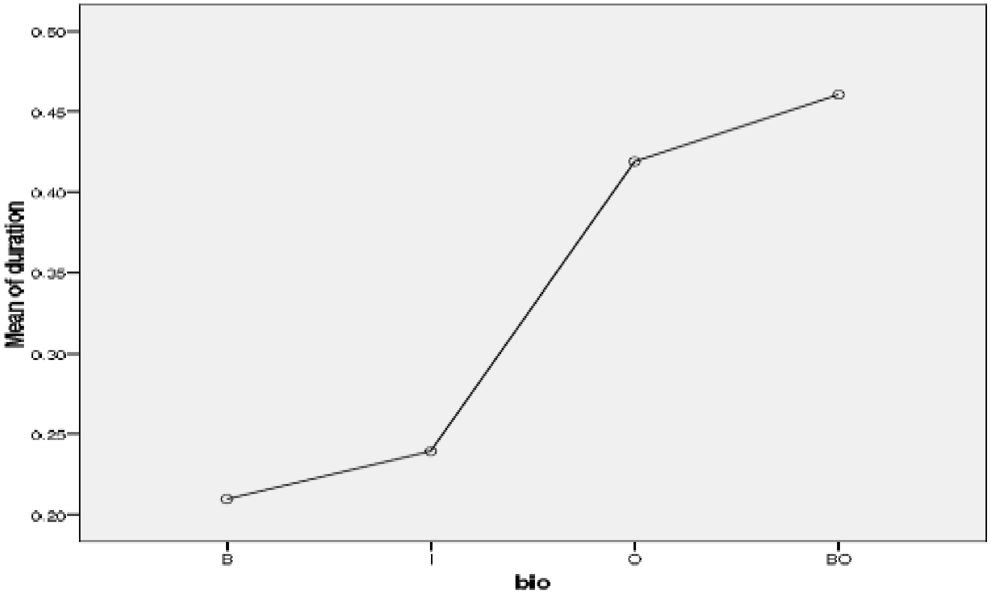
Automatic segmentation: labeling

Speaker	Token	Encoding	Function
A	because	B	Task:Inform Justify
A	twenty	I	Task:Inform Justify
A	five	I	Task:Inform Justify
A	euros	I	Task:Inform Justify
A	for	I	Task:Inform Justify
A	a	I	Task:Inform Justify
A	remote	I	Task:Inform Justify
A	how	B	Task:Set-Question
A	much	I	Task:Set-Question
A	that	I	Task:Set-Question
A	locally	I	Task:Set-Question
A	in	I	Task:Set-Question
A	pounds	O	Task:Set-Question
A	is	I	Task:Inform Justify
A	too	I	Task:Inform Justify
A	much	I	Task:Inform Justify
A	money	I	Task:Inform Justify
A	to	I	Task:Inform Justify
A	buy	I	Task:Inform Justify
A	an	I	Task:Inform Justify
A	extra	I	Task:Inform Justify
A	remote	I	Task:Inform Justify
A	or	I	Task:Inform Justify
A	a	I	Task:Inform Justify
A	replacement	I	Task:Inform Justify
A	remote	O	Task:Inform Justify

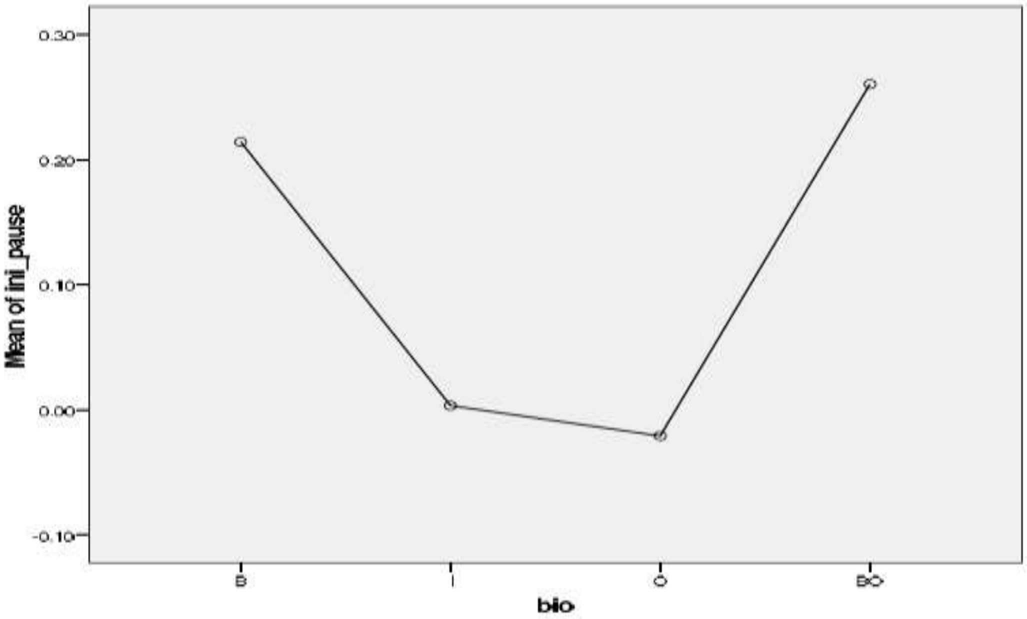
Automatic segmentation: feature selection

Feature	Pairs
duration (token)	all pairs
normalized max. pitch	O form all others
initial pause	B/I; B/O; I/BO; O/BO
normalized fraction (unvoiced/voiced)	B from others; O from others
mean pitch	all pairs except BO/B
normalized intensity	all pairs
st.dev (pitch)	all pairs
min. pitch	B/I; B/O; I/BO; BO/O
max. pitch	all except B/O
fraction (unvoiced/voiced)	all pairs
voice breaks	all pairs
intensity	all except O/BO
normalized mean pitch	I from all others
normalized st.dev (pitch)	O from all others
normalized min. pitch	all pairs
speaking rate	all pairs except BO/O

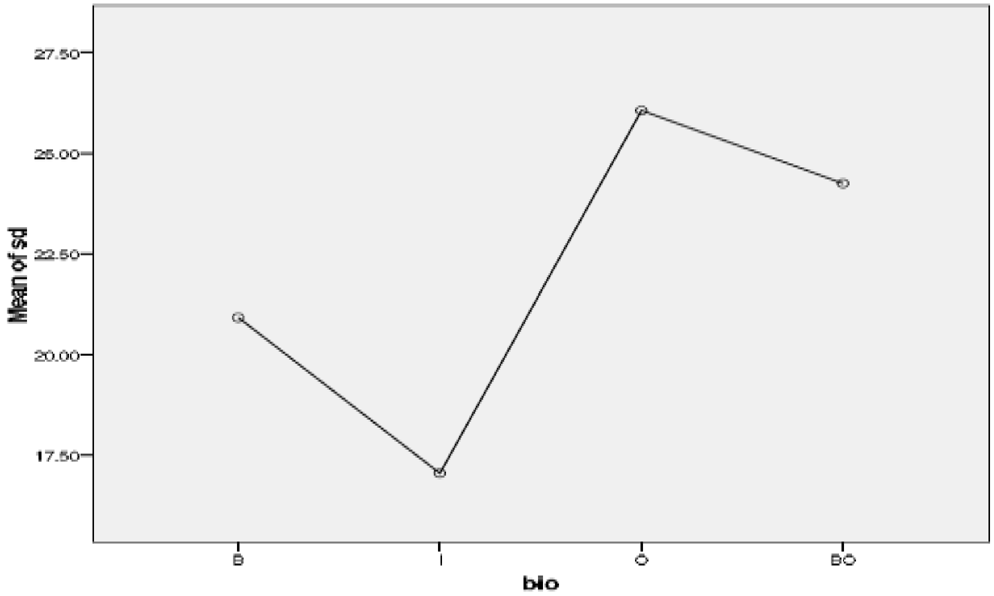
Automatic segmentation: feature selection



Automatic segmentation: feature selection



Automatic segmentation: feature selection



Automatic segmentation: classifiers

- Probabilistic, e.g. Naïve Bayes, SVM
- Rule-inducers, e.g. RIPPER
- Memory-based, e.g. IB1
- Deep learning, e.g. RNN

10-fold cross-validation (stratified)

Automatic segmentation: results

Features	Accuracy (in %)	Precision	Recall	F-scores
Prosody	79.7	0.6	0.4	0.55
Prosody + Wording	81.2	0.7	0.49	0.54
Prosody + Wording + Speaker switch	85.8	0.73	0.62	0.64
Best selected features	86.2	0.78	0.64	0.69

Begin of segment	Inside of segment	End of segment	One-token segment	Classified as
845	301	2	229	Begin of segment
74	12500	112	40	Inside of segment
1	1155	205	15	End of segment
296	149	10	1403	One-token segment

Automatic segmentation: conclusions

- Machine-learning techniques performs well
- Segment boundaries are well detectable
- But

  do we need these two steps: (1) segmentation (2) DA classification?

Answer is not necessary

DA classification as task

- Dialogue act recognition
  - A task defined by almost all dialogue modelling approaches
  - A module in almost all dialogue systems, e.g. intend in Viv
- Dialogue annotated resources: AMI, MapTask, Switchboard, etc.

DA classification as task

- Various machine learning techniques applied
  - Transformation-based learning achieved an average tagging accuracy of 75.12% for the Verbmobil corpus (Samuel et al. , 1998)
  - Hidden Markov Models (HMM) achieving a tagging accuracy of 71% on the Switchboard corpus (Stolcke et al., 2000)
  - Bayesian Networks with an average accuracy of 78% on the SCHISMA corpus (Keizer, 2003)
  - Memory-based approach (knn-classifier) with an accuracy of 73.8% on the OVIS data (Lendvai et al., 2004)
  - Neural Networks
- Various information sources are used: n-gram models or cue-phrases, syntactic and semantic features, prosodic features and context

DA classification

- Data, features and classifiers

- Labeling in multiple dimensions: dialogue act labels according to ISO 24617-2

- Tags distribution

DA classification

DA type	AMI	HCRC MapTask	SWBD	Metalogue
Commissives	2.0	21.0	3.0	19.5
Directives	8.0	15.1	13.0	20.0
Inform	26.6	11.5	36.0	20.5
Question	3.4	17.0	4.0	20.0
Other tag	60.0	35.4	44.0	20.0

Joint DA segmentation and classification:
results

Classification task	BL	NBayes	Ripper	IB1
Dimension tag	38.0	69.5	**72.8**	50.4
Task management	66.8	71.2	**72.3**	53.6
Auto-Feedback	77.9	86.0	**89.7**	85.9
Turn initial	93.2	92.9	93.2	88.0
Turn closing	58.9	85.1	**91.1**	69.6
Time management	69.7	99.2	**99.4**	99.5
OCM	89.6	90.0	**94.1**	85.6
Functional tag	25.7	48.0	**50.2**	38.9

← 97%

← 93%

Conclusions: segmentation&classification

- spoken dialogue can be described more accurately by using per-dimension segmentation instead of a single segmentation
- automatic segmentation into functional segments can be done successfully
- nevertheless, segmentation step can be avoded
- classification of DAs of functional segments for the tagset used can be done successfully in data-oriented way

Incremental classification

- human language understander does not wait trying to understand what he is reading/hearing until he has come to the end of the sentence
- evidence that human understanders construct syntactic, semantic, and pragmatic hypotheses on the fly
- not all semantic and pragmatic phenomena can be resolved incrementally
- what is the size and nature of an increment

Annotations

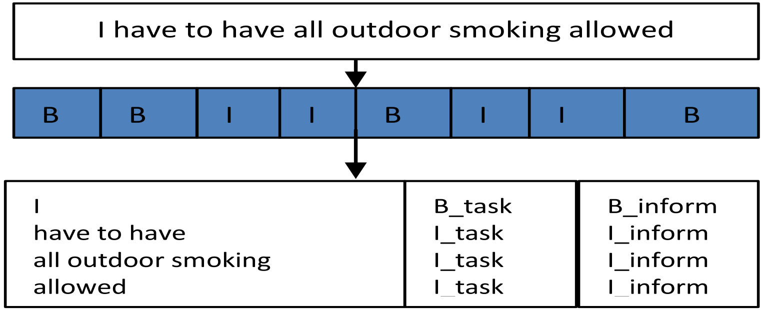
ISO 24617-2 dialogue act taxonomy

ISO 24617-2 dimension	Relative frequency (in %)
Task	47.6
AutoFeedback	18.7
AlloFeedback	2.3
Turn Management	6.6
Time Management	6.6
Discourse Structuring	14.9
Own Communication Management	2.1
Partner Communication Management	na
Social Obligation Management	1.2

*5,781 functional segments (45,479 tokens)

Data encoding

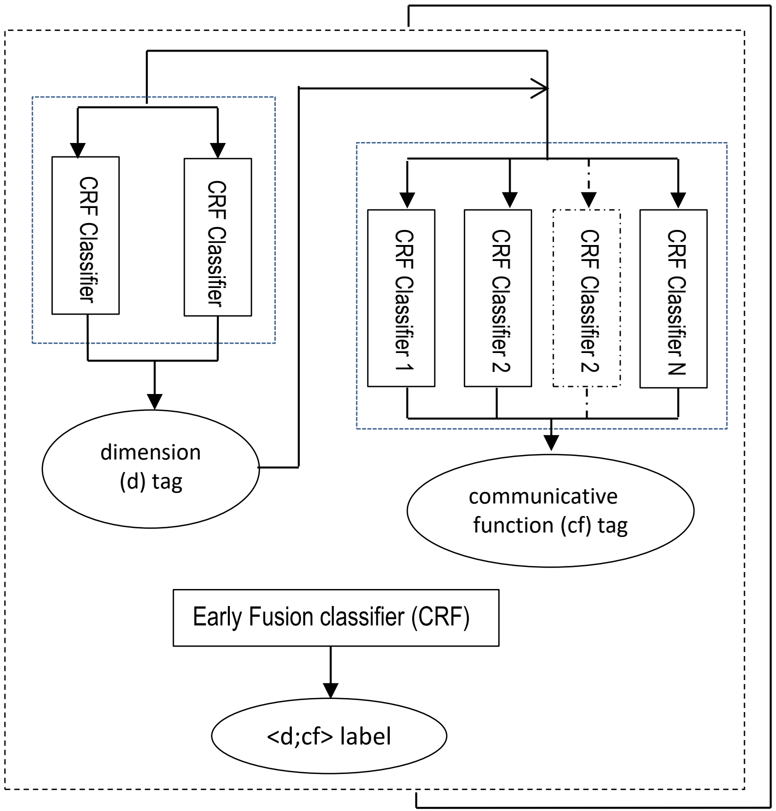
Tokens
Syntactic chunks (constituents)
Semantic chunks (entities of event and participants types, roughly semantic roles)
Prosodic chunks (inter-pausal units separated by 200ms silences coming from ASR; energy-based silence identification)

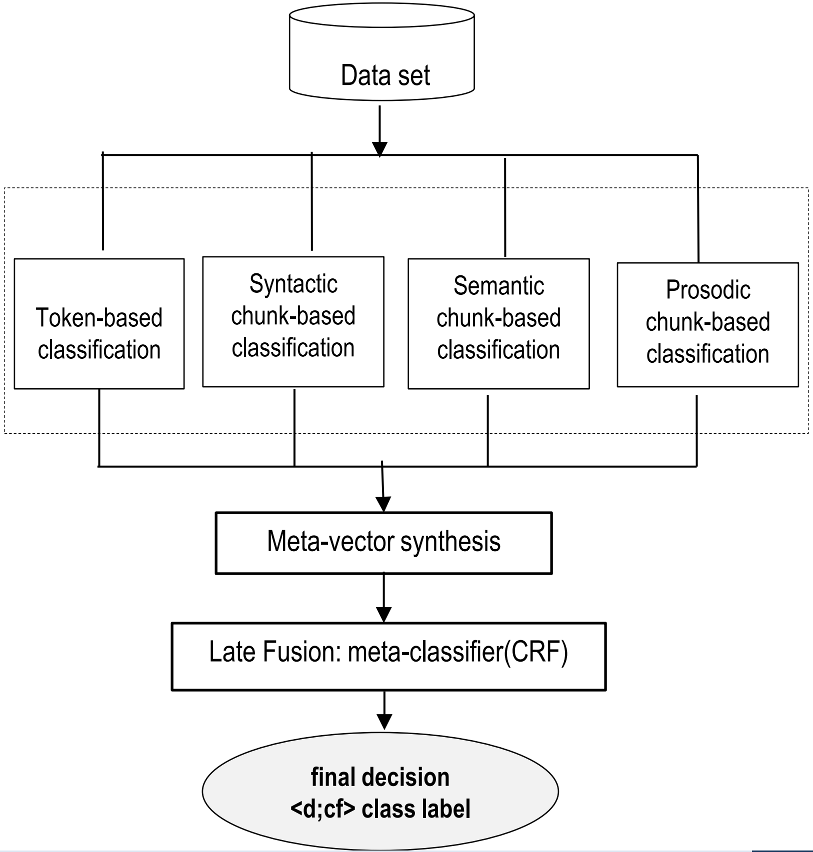


Experimental design

- Series of experiments assessing
  - Different increment types
  - Hierarchical vs cascade vs independent classification procedures
  - Features: bow, (skip) n-grams, POS tags, chunk information
  - 2 settings: simulated vs real
  - Early and late fusion steps
- Classifiers:
  - Conditional Random Fields
    - Sequence classification
    - Partial input hypotheses
    - Final complete segment hypothesis

Classification





Meta-vector synthesis

Locally computed utterance features: include tokens POS and all n-gram based features

<d> and <cf> labels sequences predicted from local features

local features used in chunk based classification

prediction space created from the output of different classifiers

token	POS	prev_token	prev_POS	skipgram	<d>	<cf>	synCHk	prosCHK	Token-based predictions	Syntactic chunk -based predictions	Semantic chunk -based predictions	Prosodic chunk -based predictions
what	PRP	null	null	null	B_task	B_Question	B_NP	B_segment	<B_task;Question>	<B_task;Question>	<B_task;Question>	<B_task;suggest>
have	ADV	what	PRP	null	I_task	I_Question	I_NP	I_segment	<I_task;Question>	<I_task;Question>	<I_task;Question>	<I_task;suggest>

Results

Setting	Simulated							Real						
	cascade		EF	hierarchical		EF	JC	cascade		EF	hierarchical		EF	JC
Task	d	cf	<d;cf>	d	cf	<d;cf >	<d;cf>	d	cf	<d;cf>	d	cf	<d;cf>	<d;cf >
Token-based	0.98	0.81	0.80	0.97	0.80	0.80	0.79	0.99	0.79	**0.77**	0.96	0.77	0.71	0.70
Chunk-based (syntactic)	0.98	0.85	**0.84**	0.96	0.83	0.82	0.80	0.98	0.78	0.70	0.96	0.74	0.64	0.69
Chunk-based (semantic)	0.98	0.84	**0.84**	0.96	0.82	0.82	0.80	0.98	0.75	0.70	0.95	0.74	0.65	0.69
Chunk-based (prosodic)	na							0.98	0.75	0.72	0.94	0.73	0.66	0.66
LF: Majority Voting	na		0.85	Na		0.82	0.79			0.78	Na		0.76	0.72
LF: Meta-classification	na		**0.85**	Na		0.82	0.80			0.80	Na		**0.77**	0.72

Conclusions: incremental processing

- Incremental dialogue recognition has the advantage that an utterance is already nearly understood even before the last token is processed.
- We have presented a machine learning based approach to incremental dialogue act classification with meta-classification approach where meta-features are synthesized from local classifiers.
- Our syntactic and semantic chunk based incremental classification produce similar results while outperforming the token based approach for manually transcribed utterances.
- Token based approach is shown to be more robust with ASR transcribed utterances