# Meeting understanding based on surface annotations

**Andrei Popescu-Belis**

**ISSCO / TIM / ETI**
**University of Geneva**
**Switzerland**

joint work with
Susan Armstrong, Alexander Clark, Maria Georgescul
Denis Lalanne, Agnes Lisowska, Sandrine Zufferey, et al.

## Institutional support

- **(IM)2** project – a Swiss NCCR
  - Interactive Multimodal Information Management
- **University of Geneva, ETI**
  - School of translation and interpreting

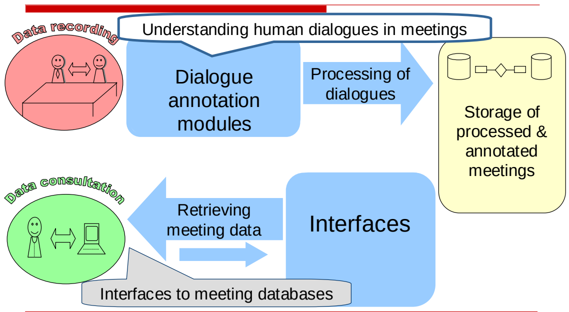


## Research on meeting processing

- Dialogue "understanding" by computers has promising applications
  - enriched meeting transcription
  - meeting summarization
  - intelligent meeting browsing
  - digital assistants for meeting rooms
  - applications to human-computer dialogue
- Desirable:
  *Fully automated minute writing application*
- Reasonable hope:
  "*Were there any questions about section 2 of the report?*"



## Meeting processing and retrieval in (IM)2






## Plan of the talk

- Introduction
- Shallow Dialogue Annotation (SDA)
  - Segmentation into episodes
  - Recognition of dialogue acts
  - Resolution of references to documents
  - Detection of discourse markers
- Use of SDA in a meeting browser
- Discussion
  - machine learning (or not) for SDA
  - cycle of evaluation-driven language processing



## Constraints on our study of dialogue processing

- Theoretical grounding
  - availability of models of the phenomenon
  - domains
    - semantics + discourse studies + pragmatics
- Application requirements
  - what users want to retrieve: analysis of user queries
  - relevance to other applications in the field
- Empirical validity
  - definitions based on examples occurring in a given corpus
  - human annotators find consistent results
- Availability of data
- Apparent feasibility

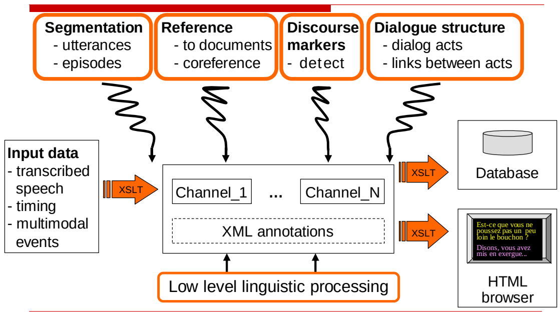
## Selected phenomena: SDA Shallow Dialogue Annotation

- Input data: timed transcript for each speaker (i.e. channel)

| | Name | Type of annotation | Scope |
|---|---|---|---|
| EP | **episodes (1)** | temporal boundaries | cross-channel |
| TO | topics/keywords | labels on EP (open set) | same as EP |
| UT | utterances | temporal boundaries | intra-channel |
| DA | **dialogue acts (2)** | labels on UT (DA tagset) | same as UT |
| RE | referring expressions | temporal boundaries | intra-channel |
| DE | **ref. to documents (3)** | pointers RE → DE | cross-modal |
| DM | **discourse markers (4)** | word classification | intra-channel |



## SDA overview






## Available data

| | Nb. x time | Media | Lg. | Annotation |
|---|---|---|---|---|
| ICSI-MR | 75 x 60' | A, T | EN | utterances, dialogue acts, discourse markers, episodes(30%) |
| IDIAP | 60 x 5' | A, V, T | EN | utterances, episodes |
| ISSCO | 8 x 30' | A, V, T, D | EN | ongoing: all |
| UniFr | 22 x 15' | A, V, T, D | FR | utterances, references to documents |

- Difficulty
  - no large dataset available yet with **all** SDA annotations



## 1. Thematic episodes: topic boundary detection [M. Georgescul]

- Goal
  - segment each meeting into coherent blocks defined by a common topic
- Methods
  - use word distribution to identify cohesive units
    - latent semantic analysis (LSA, PLSA)
  - integrate multi-word expressions
  - use discourse features (with SVM)
    - syntactic cues, speaker change, discourse markers (e.g., *well*, now), silences



## Results on topic boundary detection

- Results ($P_k$ score, ~ error rate)

| Algorithm | "Real" data | "Artificial" data |
|---|---|---|
| Baseline | 38% | 47% |
| LSA | 35% | 34% |
| C99 | 43% | 10% |

  - results on *artificial* data (merged articles) not correlated with *real* meeting data
- Next: topic characterization
  - experiments with keyword extraction vs. concept identification (EDR)



## 2. DA recognition [Clark & Popescu-Belis]

- Dialogue act
  - function of an utterance in dialogue
  - many competing theories about "function"
- DA annotation
  - presupposes segmentation of channels into utterances
  - some state-of-the-art statistical recognition methods
  - dependence on the DA tagset

## Choosing the right DA tagset

- DAMSL: independent dimensions
  - Communicative Status, Information Level, Forward Looking Function, Backward Looking Function
- SWBD-DAMSL:
  - 220 observed DAMSL labels → clustered into 42 mutually-exclusive tags
  - Statement 36%, Acknowledgement/ Backchannel 19%, Opinion 13%, Agree/Accept 5%
- ICSI-MRDA: combine (again) SWBD-DAMSL
  - ca. 7 million possible labels
- MALTUS



## MALTUS: an IM2 proposal

- Multidimensional Abstract Layered Tagset for UtteranceS
  - reduce dimensionality of ICSI-MRDA
- Structure of a MALTUS label: tags
  - main function
    - statement, question, backchannel, floor holder/grabber
  - secondary function
    - response (positive, negative or undecided), attention-related, command (performative), politeness mark, restated info.
- Number of possible labels: 770
- Conversion of ICSI-MR tags to MALTUS
  - 113,000 utterances → 50 MALTUS tags (without D)
  - more analysis and data needed to find which tags are mutually exclusive



## DA tagging in IM2 [Alex Clark]

- Objectives
  - find dimensions of MALTUS that are most easily predictable from data
  - find dependencies among tags
- Features
  - lexical (words) + contextual (surrounding tags)
- Results
  - Four way classifier (S | Q | B | H)
    - 84.9% accuracy vs. 64.1% baseline
  - Full MALTUS classifier (without "disruptions")
    - 73.2% accuracy vs. 41.9% baseline (S tag)
  - MALTUS with six classifiers trained separately
    - Primary classifier: S | H | Q | B
    - 5 secondary classifiers: PO | not PO, AT | not AT, etc.
    - 70.5% accuracy only
- Conclusion
  - separate cls. < combined cls. → dependencies between DAs



## 3. References to documents

[Lalanne & Popescu-Belis]

- Cross-media link between
  - what is said: referring expressions
  - documents and elements to which the REs refer




## Ref2doc annotation

- DIVA/University of Fribourg
  - press-review meetings (~15' each)
  - 22 meetings, 30 documents
- Ground truth annotation for training and evaluation
  - dialogue transcription, document structuring (XML)
  - RE annotation: 427 REs
  - ref2doc annotation
- Inter-annotator agreement
  - 3 annotators on 1/3 of the data
  - before discussion → after discussion
    - 96% → 100% for document assignment (3→0 errors)
    - 90% → 97% for document elements ass. (9→3 errors)



## Ref2doc algorithm based on anaphora tracking

- Loop through REs in chronological order
  - store <current document> and <current document element>
- Document assignment
  - if RE includes newspaper name → refers to that newspaper
    - <current document> set to that newspaper
  - otherwise (anaphor) → refers to <current document>
- Document element assignment
  - if RE is anaphoric → refers to <current document element>
  - otherwise → best matching document element
    - (words of RE + context) ←{*match*}→ words of document
    - <current document element> set to that element

## Results and optimization

- Best results (322 REs)
  - RE → document: 93% vs. 50% baseline (most frequent)
  - RE → doc. element: 73% vs. 18% baseline (main article)
- Optimization of features and their relevance
  - contextual features
    - only right context of the RE must be considered for matching
    - optimal size of context: ~10 words
    - relevance: when removed, ~40% accuracy only
  - (local) optimal weights for matching
    - RE ↔ title of article ≈15
      right context word ↔ title ≈10
      * ↔ content word of article ≈1
  - anaphora tracking
    - relevance: when removed, ~65% accuracy only



## 4. Discourse markers (DM)

[Zufferey & Popescu-Belis]

- Importance of DM identification
  - increase accuracy of POS tagging
  - prelude to syntactic analysis
  - indicate global discourse structure
  - indicate coherence relations (à la RST) between utterances
  - serve as features for the automatic detection of dialog acts
- Two markers were studied
  - “like” - signals approximation
  - “well” - marks topic shift, or correction
- Problem
  - both lexical items are ambiguous: they can function as a discourse marker or as something else (e.g., verb or adverb)
  - need to **disambiguate occurrences: DM vs. non-DM**



## Examples

1a. It allows you to enter things ***well***.

1b. So they'll say ***well*** these are the things I want to do.

2a. Did you ***like*** the movie?

2b. Most of our meetings are uh meetings currently with ***like***, five, six, seven, or eight people.

- How to detect only “pragmatic” uses? → (b) *vs.* (a)



## Disambiguation of DM *like* by humans using prosodic cues

- 1st experiment: only with transcript
- 2nd experiment: transcript linked to audio
- Annotators had to classify each occurrence of *like* as DM or non-DM
- Inter-annotator agreement
  - $\kappa = 0.74$ (> 0.67)
  - reliable task
  - prosodic cues are crucial



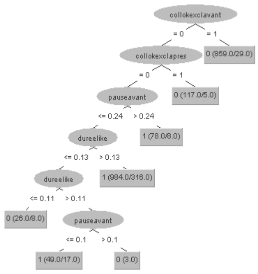
## Statistical training of DM classifiers

- Decision trees + C4.5 training (Quinlan / WEKA)
- Features characterizing DM vs. non-DM uses
  - “negative” or excluding collocations
  - duration of item
  - duration of pause before *like*
  - duration of pause after *like*
- Set of positive and negative examples from ICSI-MR
  - ~4500 for *like* and ~4100 for *well*
- Results of the training
  - binary decision tree classifier (DM / non-DM)
  - measure of the discrimination power: 10 times cross-validation



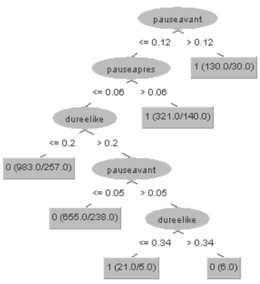
## Results for DM classification

- Scores for *like*: best classifier
  **r = 0.95 / p = 0.68 / $\kappa$ = 0.65**
- Conclusions
  1. Importance of collocation filters
  2. A pause before *like* indicates a DM in 91% of the remaining cases
  3. Other factors are relevant too, but quite redundant
     → prosody



## Without collocation filters

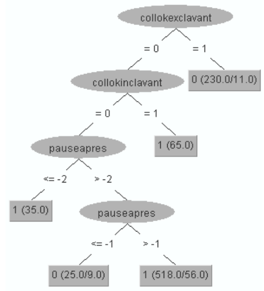
- Scores of best classifier
  - r = 0.35 / p = 0.6 / $\kappa$ = **0.23**
- Conclusions
  1. Other features are relevant too
  2. Best temporal feature: a pause before or after *like*
  3. Temporal features are redundant when collocations can be used
- Prosody is relevant to human annotators
  → try to find other relevant prosodic features






## Best classifier for *well* as a DM

- Scores
  - r = 0.97 / p = 0.91 / $\kappa$ = **0.81**
- Conclusions :
  1. Importance of collocations
  2. A pause after *well* indicates the presence of a DM
- Use of collocations only
  - r = 0.98 / p = 0.89 / $\kappa$ = 0.78
- Relevance of other features?
- Use of "pause after" only
  - r = 0.96 / p = 0.77 / $\kappa$ = 0.45

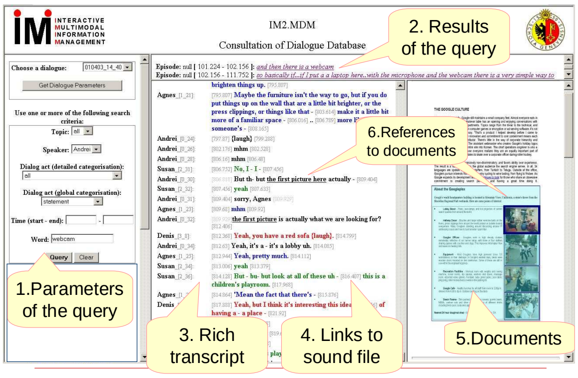





## Snapshot / Demo

Use of SDA in a meeting browser

## TQB: Transcript-based query & browsing interface






## Summary: machine learning techniques and their scores

| | Tag set | Method | Baseline | Accuracy |
|---|---|---|---|---|
| DA | MALTUS | MaxEnt | ~40% | 70-73% |
| EP | Boundaries | LSA/C99 | 67% | 60-(90)% |
| DE | RE→DE | Rule-based | ~20% | 73% |
| DM | DM/non-DM | Decision trees, C4.5 | 36% (*like*)<br>66% (*well*) | 81%<br>91% |

- Machine learning appears to be relevant to semantic/pragmatic annotations
- More or less transparent statistical models



## SDA: machine learning or not?

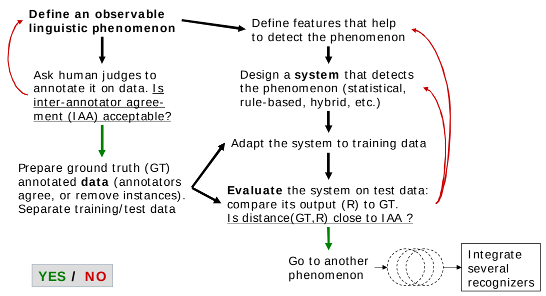
- Use of machine learning when…
  - enough annotated data for training
  - enough low-level relevant features
  - unknown optimal relations between features and annotations
  - **DA, EP, (TO), DM**
    - possibility to add some obvious hand-crafted rules
- Use of hand-crafted rules or classifiers when…
  - not enough data to learn relations between features and annotations
  - **(UT), (RE), RE→DE**
    - possibility to optimize automatically the hand-crafted rules
- Possibilities to use a mix them

# Future work

- Integration: "multi-agent dialogue parser"
  - each module generates annotations
  - loop through modules until no annotation can be added
- Extensions
  - add new modules, improve existing ones: TO, RE, ...
  - use multimodal features: prosody, face expression, ...
- Relevance of SDA annotations to meeting browsing
  - design interfaces to annotated database
  - test them with/without access to annotations



# Conclusion: The basis of evaluation-driven language processing






# References


- Clark A. & Popescu-Belis A. (2004) - Multi-level Dialogue Act Tags. In *Proc. SIGDIAL'04*, Cambridge, MA, p.163-170.
- Lisowska A., Popescu-Belis A. & Armstrong S. (2004) - User Query Analysis for the Specification and Evaluation of a Dialogue Processing and Retrieval System. In *Proc. LREC 2004*, Lisbon, p.993-996.
- Popescu-Belis A., Clark A., Georgescul M., Zufferey S. & Lalanne D. (2005) - Shallow Dialogue Processing Using Machine Learning Algorithms (or not). In Bengio S. & Bourlard H., eds., *Machine Learning for Multimodal Interaction*, LNCS 3361, Springer-Verlag, Berlin, p.277-290.
- Popescu-Belis A. & Lalanne D. (2004) - Ref2doc: Reference Resolution over a Restricted Domain. In *Proc. ACL 2004 Workshop on Reference Resolution and its Applications*, Barcelona.
- Zufferey S. & Popescu-Belis A. (2004) - Towards Automatic Disambiguation of Discourse Markers: the Case of 'Like'. In *Proc. SIGDIAL'04*, Cambridge, MA, p.63-71.