Shallow Dialogue Annotation (SDA) for a Transcript-based Query and Browsing interface (TQB) to meetings

Andrei Popescu-Belis

Alexander Clark
Maria Georgescul
Sandrine Zufferey
ISSCO/TIM/ETI
University of Geneva

Denis Lalanne
DIVA/DIUF
University of Fribourg

Research on meeting processing

- Desirable:

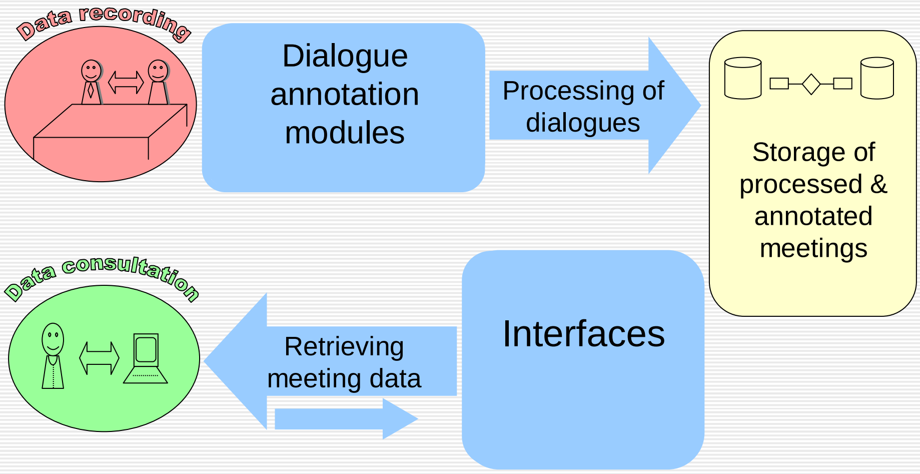
  Fully automated minute writing application.

- Reasonable hope:

  "*Were there any questions about section 2 of the report?*"

- Range of applications
  - enriched meeting transcription
  - meeting summarization
  - intelligent meeting browsing
  - digital assistants for meeting rooms

Meeting processing & retrieval



Constraints on our study of dialogue processing

- Theoretical grounding
  - availability of models of the phenomenon
  - known active research topics
- 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 of individual channels

	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 doc. elements (3)	pointers RE → DE	cross-modal
DM	discourse markers (4)	word classification	intra-channel

1. Thematic episodes (EP)

- Segmentation of meeting into coherent blocks defined by a common topic
- Representation of input
  - based on lexical items
  - vector space model
- Training
  - use of latent semantic analysis (LSA) to reduce dimensionality of word/frequency matrix
  - singular value decomposition
  - deletion of smallest diagonal terms

Application of LSA

- Test phase
  - thematic distance between consecutive utterances
  - computed by projection on the reduced lexical space
  - segmentation at lowest points
- Evaluation
  - various conditions on expected number of boundaries
  - various scoring methods
- Results
  - better to train and test on same type of data
    - e.g., texts from Brown corpus, TDT data
    - 10-20% error rate
  - ICSI-MR data
    - 35% error rate
  - also used: C99

2. DA recognition

- Dialogue act: function of an utterance in dialogue
  - presupposes segmentation of channels into utterances
- Many tag sets available, e.g. ICSI-MRDA (~7.10^6 labels)
- **MALTUS** (~500 labels)
  - main function
    - statement, question, backchannel, floor holder/grabber
  - secondary function
    - response (positive, negative or undecided), attention-related, command (performative), politeness mark, restated information
- Dataset
  - conversion of ICSI-MR to MALTUS
  - 50 occurring MALTUS labels in **113,560 utterances**

DA tagging: objectives and features

- Our objectives
  - find dimensions of MALTUS that are most easily predictable from data
  - find hidden dependencies among tags
  - different from tagging using a language model ([Stolcke et al. 2000])
    - features: word n-grams + dialogue model (sequence of DAs)
- Simplifying assumption
  - allow access to the gold standard DA of surrounding utterances
  - ➢ use maximum entropy classifier (no decoding)
- Features
  - lexical
    - 1000 most frequent words + their positions in utterance
  - contextual
    - label of the preceding utterance in the same channel & in different channels (x2)
    - label of utterances overlapping with current one & contained in the current one (x2)

DA tagging: results

- Four way classifier (S | Q | B | H)
  - 84.9% accuracy vs. 64.1% baseline
- Six way classifier (S | Q | B | H | disruption | indecipherable)
  - 77.9% accuracy vs. 54.0% baseline
- **Full MALTUS classifier** (but no disruptions)
  - 73.2% accuracy vs. 41.9% baseline (S tag)
- MALTUS with six classifiers trained separately
    - Primary classifier: S | H | Q | B
    - Five secondary classifiers: PO | not PO, AT | not AT, etc.
  - 70.5% accuracy only

3. References to documents (RE→DE)

- Cross-media link between
  - what is said: referring expressions
  - documents and elements referred to
- Pre-requisites
  - detection of referring expressions (RE)
    - ongoing work
  - automatic detection of document elements = document structuring

Ref2doc annotation

```
<dialog>
 <channel id="1" name="Denis">
   ...
   <er id="12">The title</er>suggests that the issue…
 </channel>
 ...
 <ref2doc>
 ...
 <ref er-id="12"
      doc-file="LeMonde030404.Logic.xml"
      doc-id="//Article[@ID='3']/Title"/>
     ...
 </ref2doc>
</dialog>
```

Referring expression (uttered by Denis)

Document referred to (XML logical structure)

Document element (XPath)

Algorithm based on anaphora tracking (hand-crafted)

- 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)

- Useful to detect
  - 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 [Zufferey & Popescu-Belis 04]
  - "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**

Statistical training of DM classifiers

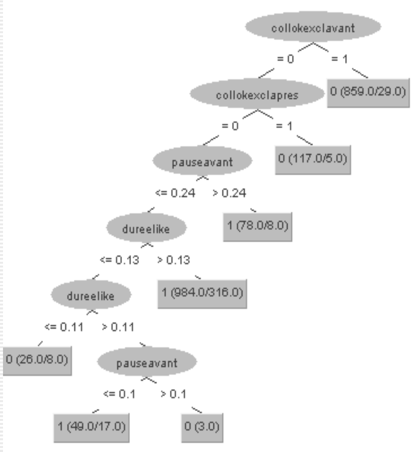
- Decision trees + C4.5 training (Quinlan / WEKA)
- Features characterizing DM vs. non-DM uses
  - "negative" or excluding collocations
  - duration of item
  - duration of pause before *like*
  - duration of pause after *like*
- Set of positive and negative examples from ICSI-MR
  - ~4500 for *like* and ~4500 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 / κ = 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

- Scores for *well*: best classifier
  - **r = 0.97 / p = 0.91 / κ = 0.81**



26/11/2004 Shallow Dialogue Annotation for TQB 17

Synthesis: machine learning for SDA

- 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

26/11/2004 Shallow Dialogue Annotation for TQB 18

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*) 66% (*well*)	81% 91%

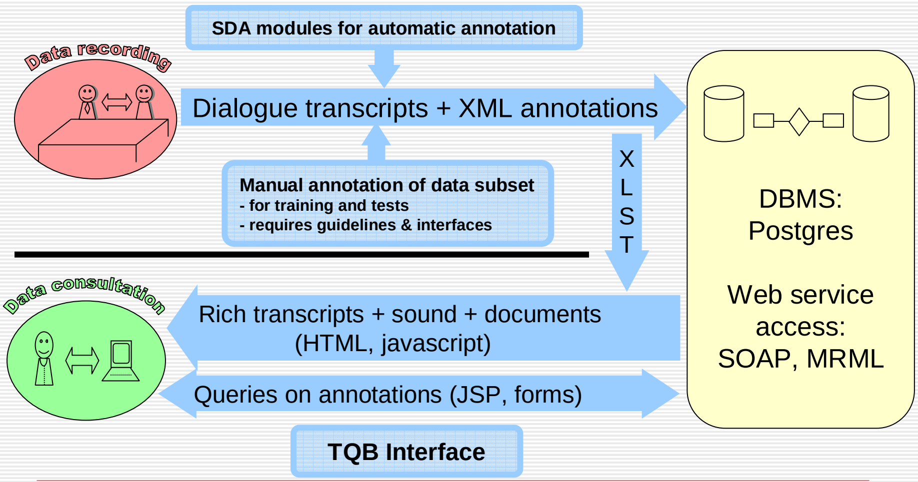
- Machine learning appears to be relevant to both semantic and pragmatic annotations
- More or less transparent models

Available data

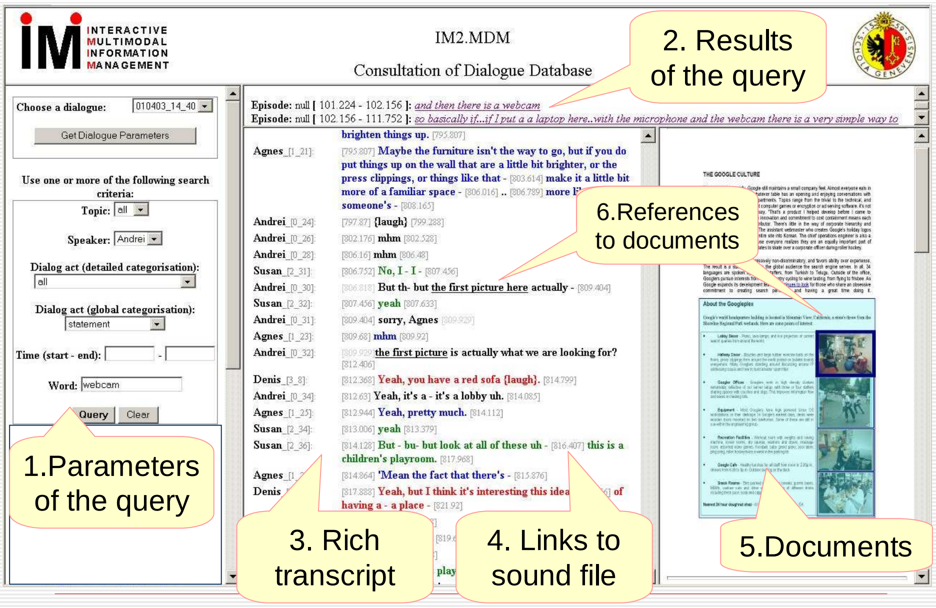
	Nb. x time	Media	Lg.	Annotation
ICSI-MR	75 x 60'	A, T	EN	utterances, dialogue acts, discourse markers, episodes(30%)
IDIAP	60 x 5'	A, V, T	EN	utterances, episodes
ISSCO	8 x 30'	A, V, T, D	EN	ongoing: all
UniFr	22 x 15'	A, V, T, D	FR	utterances, references to documents

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

Data formats and tools



TQB: Transcript-based query & browsing interface



Future work

- Integration of SDA modules
  - each module generates annotations based on features and other existing annotations
  - trigger modules in a loop until no annotation can be added
- Extensions
  - improve/extend existing modules: TO, RE, ...
  - add new annotation modules: depending on user needs
  - make use of new features, especially from other modalities: prosody, face expression, ...