UiO : University of Oslo

Selected Topics in Spoken Dialogue Processing

Pierre Lison
University of Oslo, Dep. of Informatics

Språkteknologisk seminar @ LNS
November 16, 2011

Introduction

- Presentation of a series of research challenges
  - Common denominator: *spoken dialogue processing*
  - Descriptive and computational perspective
- Objectives:
  - convince you that spoken dialogue offers interesting, unexplored challenges for NLP
  - motivate you to do research with me on some of these issues ;-)

Introduction (2)

- 4 «open questions» that could serve as starting points for further research
    - side-projects from my Ph.D. work
- Acknowledgements:
    - recorded samples from «*Norske talespråkskorpus - Oslo delen*» (NoTa), collected and annotated by our colleagues at the Tekstlaboratoriet
    - Timo Baumann (Uni. Hamburg) for his comments

[http://www.tekstlab.uio.no/nota/oslo/index.html]

Outline of the talk

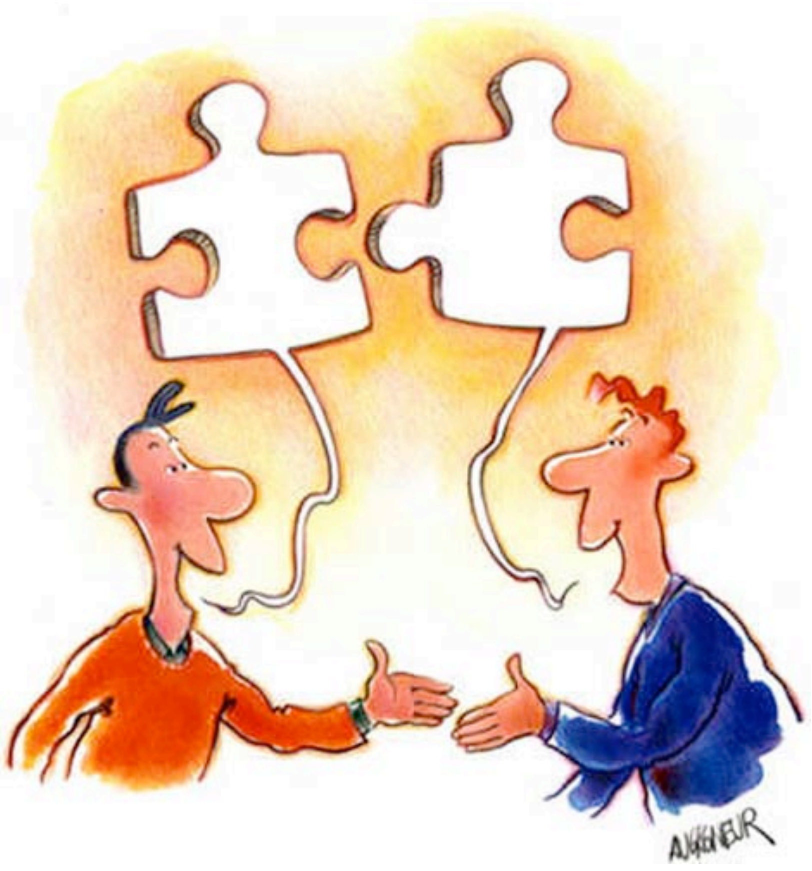
- Generalities about dialogue
- Selected topics:
    - Incremental understanding
    - Adaptive feedback generation
    - Treatment of disfluencies
- Conclusion

Outline of the talk

- **Generalities about dialogue**
- Selected topics:
  - Incremental understanding
  - Adaptive feedback generation
  - Treatment of disfluencies
- Conclusion

What is dialogue?

- Spoken ("verbal") + possibly non-verbal interaction between two or more participants
- Dialogue is a joint, social *activity*, serving one or several purposes for the participants
- What does it mean to view dialogue as a **joint activity**?



Dialogue as joint **activity**

- Each utterance is an *action* performed by the speaker
  - Types of dialogue acts: promising, ordering, warning, asking, replying, maintaining social contact, etc.
- «*Language as action*» perspective
- Dialogue acts exhibit both
  - an *internal* structure (arguments, adjuncts, etc.)
  - an *external* structure (rhetorical relations, references, etc.)

[John Searle (1969), «*Speech acts*», CUP]

Turn-taking

- Dialogue participants takes *turns*
    - Turn = continuous contribution from one speaker
- How are turns taken and released?
    - Verbal/non-verbal cues + social conventions
- Surprisingly fluid in normal conversations:
    - less than 5 % overlap
    - Minimal pauses between speakers (<100ms)

[Duncan (1972): «Some Signals and Rules for Taking Speaking Turns in Conversations», in *Journal of Personality and Social Psychology*]

Example of turn-taking

Speaker 1: han vil bo i skogen ?

Speaker 2: # altså hvis jeg hadde kommet og sagt " skal vi flytte i skogen ? " så hadde han sagt ja

Speaker 1: mm

Speaker 2: men jeg vil ikke bo i skogen

Speaker 1: nei det skjønner jeg

Speaker 2: så vi må jo finne et sted som er mellomting og det jeg vil ikke bo utpå landet # i hvilken som helst (uforståelig) ...

Speaker 1: * men det kommer jo an på hvor i skogen da

Incrementality

- Processing of spoken dialogue is strongly *incremental*
  - Both for comprehension and production
  - Very low latency
- Continuous projection of *hypotheses* on how the interaction is likely to unfold
  - Predictive mechanisms central to human cognition
- **Downside**: speakers construct their utterances «as they go», leading to numerous disfluencies

[Van Berkum, J. J. A. (2010) in *Italian Journal of Linguistics*]

Dialogue as **joint** activity

- Dialogue is a joint, *collaborative* process between the participants
  - Cooperative responses
  - Cooperative interpretation (beyond literal meaning)
  - Taking initiative
- Importance of *grounding* to continually ensure mutual understanding
- Role of alignment and imitation (cf. previous talk)

Grounding in dialogue

- Participants establish and gradually refine their *common ground*
  - Common ground = shared knowledge
- Grounding mechanisms:
  - Backchannels, (implicit, explicit) feedbacks
  - Verifications
  - If a problem arises: *clarification* and *repair* strategies

[Clark, H. H. (1996), «*Using Language*», CUP]

Example of grounding

Speaker 1: vi vasker den hver dag vi # vi har mopp

Speaker 2: mm ## ja det er fort og faren til M27 legger nytt teppe han # det er gjort på to timer ## så det er fort gjort

Speaker 1: ja ## da er ikke noe sak

Speaker 2: vi har skifta teppe tre ganger allerede han gjør det gratis

Speaker 1: hæ ?

Speaker 2: vi har skifta teppe tre ganger og # han han ...

Speaker 1: * jeg skjønner ikke hvorfor dere har teppe

Speaker 2: jeg syns det var rart jeg òg # men e # (sibilant)

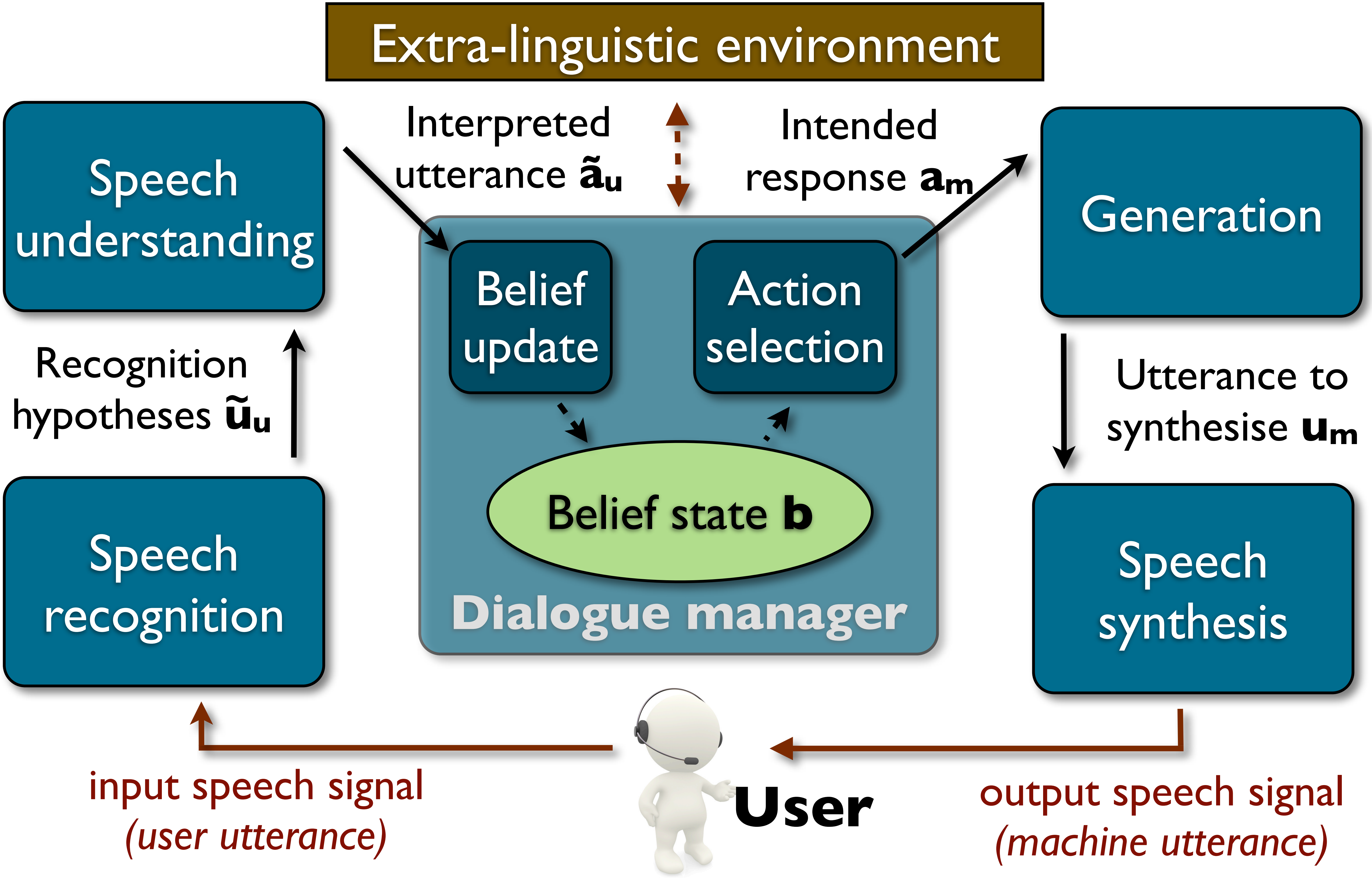
Taking stock

- Dialogue seen as a **joint activity:**
  - Dialogue acts
  - Turn-taking
  - Incrementality
  - Cooperation
  - Grounding

→

How can these insights help us design better dialogue systems?

Dialogue systems architecture



Extra-linguistic environment

Speech understanding

Interpreted utterance $\tilde{\mathbf{a}}_u$

Intended response $\mathbf{a}_m$

Generation

Belief update

Action selection

Belief state $\mathbf{b}$

Dialogue manager

Recognition hypotheses $\tilde{\mathbf{u}}_u$

Utterance to synthesise $\mathbf{u}_m$

Speech recognition

Speech synthesis

User

input speech signal *(user utterance)*

output speech signal *(machine utterance)*

Outline of the talk

- Generalities about dialogue
- **Selected topics:**
  - Incremental understanding
  - Adaptive feedback generation
  - Treatment of disfluencies
- Conclusion

Outline of the talk

- Generalities about dialogue
- Selected topics:
  - **Incremental understanding**
  - Adaptive feedback generation
  - Treatment of disfluencies
- Conclusion

Incrementality in dialogue systems

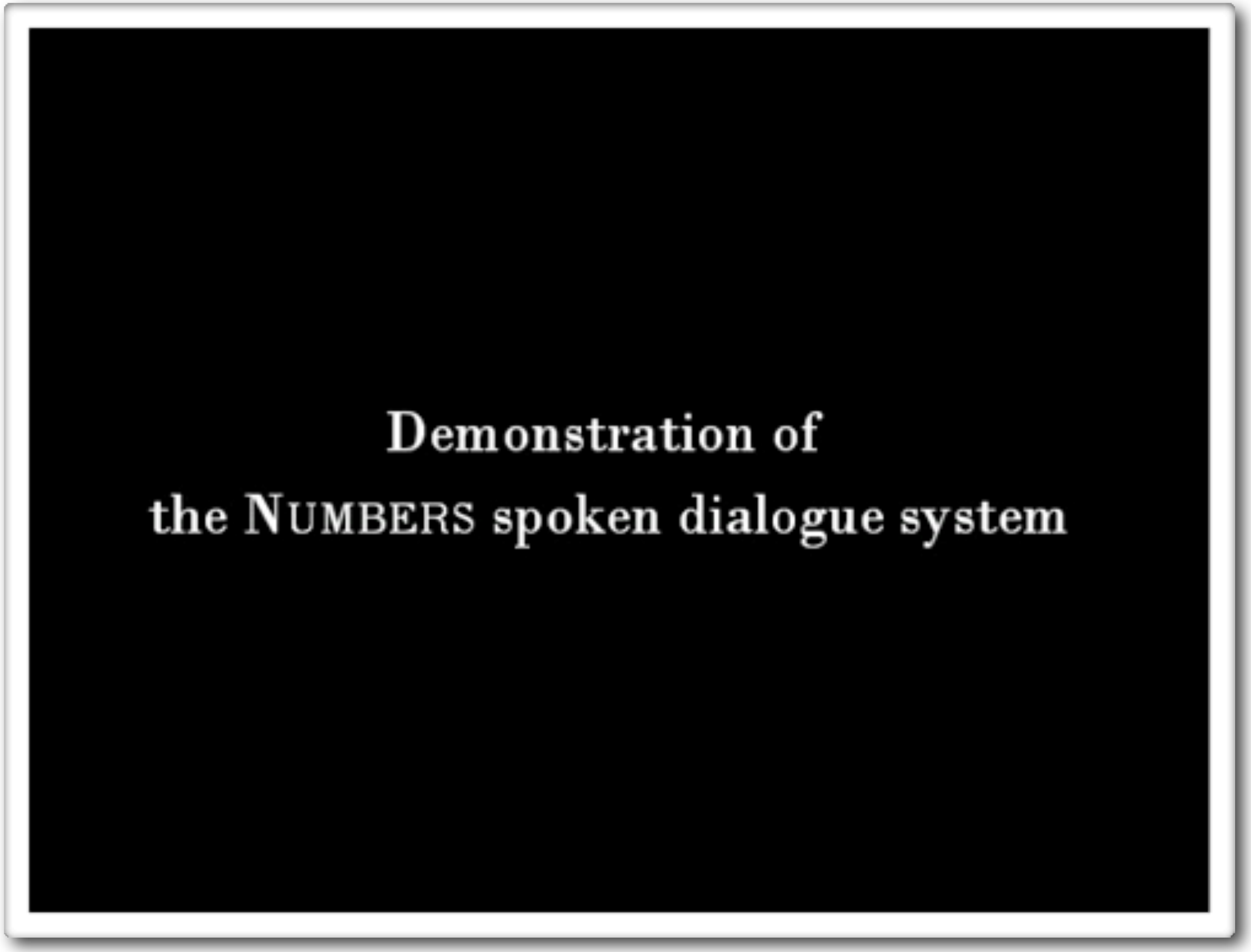
- *Incrementality* currently a hot topic in spoken dialogue systems research
- Motivation: go beyond the «ping-pong»-like behaviour of current-day systems
  - More reactive turn-taking behaviour
  - More robust & efficient interpretation
  - More responsiveness (early feedbacks, interruptions)

Incremental processing model

- David Schlangen's generic incremental model of dialogue processing:
  - Network of interconnected processes, transferring information via input and output buffers
  - Incremental Unit (IU) = basic representational unit
  - IUs are interconnected via various relations, forming a full network within & across processing levels
  - 3 basic operations on IUs: *update*, *purge* and *commit*

[Schlangen, D. and Skantze G. (2009) «A General, Abstract Model of Incremental Dialogue Processing», in Proceedings of EACL 2009.]

Example of incremental system



Demonstration of

the NUMBERS spoken dialogue system

[Skantze G. and Schlangen, D. (2009), «Incremental dialogue processing in a micro-domain», in Proceedings of EACL 2009.]

Incremental understanding

- Let’s focus on the specific problem of incremental understanding
    - Goal: extract a representation of the dialogue act from the raw recognised utterance (N-best list)
- Many systems rely on simple keyword spotting, ignoring the utterance structure
    - Alternative: extract relevant *syntactic features* with a parser, and exploit them in dialogue act recognition

Incremental parsing

- Main challenges: recognition errors, disfluencies (more on this later)
- Furthermore: incremental parsing for dialogue is not always *monotonic*
    - ASR recognition lattice at time $t+1$ is not necessarily a monotonic extension of the lattice at time t
    - But incremental parsers generally rely on a single sentence which does not change over time

Incremental understanding (2)

Open question 1: how can we extend existing algorithms for incremental parsing to:

- work on recognition lattices (with probabilities) instead of single sentences?
- handle non-monotonic inputs?

Outline of the talk

- Generalities about dialogue
- Selected topics:
  - Incremental understanding
  - **Adaptive feedback generation**
  - Treatment of disfluencies
- Conclusion

Transparency in dialogue

- We have seen that grounding acts were essential to mutual understanding
    - Clarifications, verifications, repairs, feedbacks etc.
- Often difficult for the user to know what the current system state is
    - What is the system attending to, what is already understood and what is not?
    - Dialogue system should be as *transparent* as possible

Feedback generation

- We focus here on simple *system* feedbacks
  - *Various modes*: continued attention, vocalisations, non-verbal signals, explicit or implicit responses, etc.
  - Different *levels of understanding*, from simple detection of a sign to its complete interpretation
  - *Timing* is crucial for all
- How to decide when to generate feedback, and in which form?

Machine learning approach?

- Selecting the right type of feedback depend on various factors interacting in complex ways:
  - Confidence levels & grounding in current variables
  - Global features: noise level, user type, history of previous feedbacks, etc.
- Encoding such complex strategies in handcrafted heuristics is unwieldy

Machine learning approach (2)

- Instead of heuristics, can we learn optimal strategies for feedback generation from data?
  - Supervised learning problem?
  - Potential issues: uncertain features (hidden variables), representation of timing information
- Data could be provided by recordings of Wizard-of-Oz experiments
  - Problem: limited amounts of data!

ML-based feedback generation?

Open question 2: can we apply machine learning on Wizard-of-Oz data to *learn* how to generate proper feedback?

- If yes, which features to use?
- Which learning algorithm?
- How to take uncertain variables into account?
- How to take timing into account?
- Can we show that such approach yields more transparent and adaptive behaviours?

Outline of the talk

- Generalities about dialogue
- Selected topics:
  - Incremental understanding
  - Adaptive feedback generation
  - **Treatment of disfluencies**
- Conclusion

Disfluencies in dialogue

- As we have seen, speakers construct their utterances «as they go»
  - Production leaves a *trace* in the speech stream
  - Silent and filled pauses, fragments
  - Frequent repetitions, corrections, repairs
  - *Meta-communicative* dialogue acts, where the user reflects and comments on her/his own «performance»
  - Many *non-sentential utterances* [NSUs], interpreted against the broader context of the interaction

Disfluency detection

- Can we automatically detect disfluencies?
- Influence of Shriberg's foundational work on speech disfluencies in the mid-90's
  - considered types of disfluencies: filled pauses, repetition, substitution, insertion, deletion, speech error
- Switchboard corpus often used for evaluations
  - speech corpus of telephone conversations
  - explicitly annotated with disfluencies

Shriberg's disfluency model

- Internal structure of a disfluency:

$$\text{Book a ticket } \underbrace{\text{to Boston}}_{\text{reparandum}} \; \underbrace{\text{uh I mean}}_{\text{interregnum}} \; \underbrace{\text{to Denver}}_{\text{repair}}$$

  - reparandum: part of the utterance which is edited out
  - interregnum: (optional) filler
  - repair: part meant to replace the reparandum

[Shriberg (1994), «Preliminaries to a Theory of Speech Disfluencies», Ph.D thesis, UC Berkeley]

Basic examples of disfluencies

- Repetitions

$$\text{robot now go to } \underbrace{\text{the hallway}}_{\text{reparandum}}\ \underbrace{\text{the hallway}}_{\text{repair}}$$

- Corrections:

$$\text{ok and then turn } \underbrace{\text{right}}_{\text{reparandum}}\ \underbrace{\text{no sorry I mean}}_{\text{interregnum}}\ \underbrace{\text{left}}_{\text{repair}}$$

- Rephrasing/completion:

$$\text{robot please give me } \underbrace{\text{the ball}}_{\text{reparandum}}\ \underbrace{\text{yes}}_{\text{interregnum}}\ \underbrace{\text{the red one on your left}}_{\text{repair}} \text{ exactly}$$

General remarks on disfluencies

- All parts of a disfluency may carry *meaning* relevant for interpretation
  - Even filled pauses such as «uh» and «um»
- Levelt: reparandum and repair are of syntactic types that *could* be joined by a conjunction
- Pervasive phenomena: about 6% of the words in spontaneous speech are «edited»

[Levelt W. (1983), « Monitoring and self-repair in speech», in *Cognitive Science*.]

Noisy channel approach

- Motivation: words in reparandum usually closely related to those in the repair
- Given observed sentence Y, search for:

$$\hat{X} = \operatorname*{argmax}_{X} \Pr(Y|X)\Pr(X)$$

  - Language model $\Pr(X)$: bigram, trigram, syntax-based
  - Channel model $\Pr(Y|X)$: TAG matching reparandum to repair using deletion, insertion, substitution.

[Johnson, M. & Charniak, E. «A TAG-based noisy channel model of speech repairs», Proceedings of ACL 2004]

Treatment of disfluencies

- Research effort mostly targeted on disfluency *detection* in *human-human* dialogues
- Not so much work on full disfluency *treatment* in *human-machine* dialogues
  - **Easier**: human-machine interaction is usually less disfluent (human users adapt to the machine)
  - **More difficul**t: need to work on real ASR outputs instead of gold transcripts
  - What do we do with the disfluency after detection?

Treatment of disfluencies (2)

Open question 3: how can we handle disfluencies in a *end-to-end* dialogue system?

- What is the best way to treat disfluencies *after* detection?
- How to simultaneously handle speech recognition errors and disfluencies?
- Does the treatment of disfluencies improve the system task performance?

Beyond basic disfluencies...

så gikk jeg e flytta vi til Nesøya da begynte jeg på barneskolen der

og så har jeg gått på Landøya ungdomsskole # som ligger ## rett over broa nesten # rett med Holmen

jeg gikk på Bryn e skole som lå rett ved der vi bodde den gangen e barneskole

videre på Hauger ungdomsskole

da hadde alle hele på skolen skulle liksom # spise julegrøt og det va- det var bare en mandel

og da var jeg som fikk den da ble skikkelig sånn " wow # jeg har fått den " ble så glad

Limitations

- Extension of disfluency not always clear
- Disfluencies essentially viewed as «noise» or «performance errors», outside the scope of natural language syntax
  - *But*: disfluencies are often meaningful!
  - *But*: widespread and universal phenomena
  - *But*: close similarities with other syntactic phenomena such as coordination

Paradigmatic piles

- Insights from descriptive linguistics: Claire-Blanche Benveniste's work on spoken French
- Idea of «paradigmatic piles»
  - non-functional relations between phrases (i.e. relations without head-dependent asymmetry)
  - Paradigmatic pile = position in a utterance where the "syntagmatic unfolding is interrupted", and the same syntactic position hence occupied by several linguistic objects
  - represented in a grid

[Benveniste, C.-B. (1998), «*Le francais parlé: études grammaticales*», Éd. du CNRS]

Disfluency and coordination

(a) Felix is a linguist, maybe a computer scientist [Disfl]
(b) Felix is a linguist uh maybe a computer scientist [Disfl]
(c) Felix is a linguist or maybe a computer scientist [Coord]
(d) Felix is a linguist and maybe a computer scientist. [Coord]

- (c) has the same interpretation as (b)
- (a) can either be interpreted «disjunctively» as in (b), (c), or «additively» as in (d)
- The syntactic types accepted in disfluencies and in coordination are similar (cf. Levelt's rule)

[Gerdes K., Kahane S. (2009), «Speaking in piles: Paradigmatic annotation of French spoken corpus», Processing of the 5th Corpus Linguistics Conference]

Disfluency and coordination (2)

(a)	Felix is		a linguist
		maybe	a computer scientist
(b)	Felix is		a linguist
		uh maybe	a computer scientist
(c)	Felix is		a linguist
		or maybe	a computer scientist
(d)	Felix is		a linguist
		and maybe	a computer scientist.

- Paradigmatic piles provide an unified treatment of (a)-(d)
- «maybe», «and» etc. are are *pile markers*
  - Pile structure similar for the 4 examples, but the final interpretation slightly different due to the distinct markers

Detailed example

vokst opp i et stort stort hus # med tre etasjer og (latter) ## mange rom i hver etasje og

store rom ## god plass # lun e # lun e # sånn gårdsstemning # i hvert rom ja

og ## ja ## nå bor jeg jo i en (latter) # mer urban # minimalistisk # moderne leilighet

Grid analysis of example

vokst opp i et stort					
	stort hus med			tre etasjer	
			og	mange rom i hver etasje	
			og	store rom	
				god plass	
				lun e	
				lun e	
				sånn gårdsstemning i hvert rom ja	
og					
ja					
nå bor jeg jo i en mer		urban			
		minimalistisk			
		moderne			leilighet

Paradigmatic piles: discussion

- Piles provide a descriptive account of various syntactic phenomena
  - disfluencies, reformulation, appositions, coordinations, etc.
  - Piles viewed as a *complement* to dependency relations
  - Syntax expressed as a two-dimensional structure
- Purely descriptive account: no formal definitions of the rules and constraints on the piles
- Framework used to provide detailed syntactic annotation for corpora of spoken French

Treatment of disfluencies

Open question 4: can we define a syntactic treatment of disfluencies which goes beyond the noisy channel approach?

- How would disfluencies be annotated?
- Can we train or adapt a data-driven parser to capture such constructions?

Outline of the talk

- Generalities about dialogue
- Selected topics:
  - Incremental understanding
  - Adaptive feedback generation
  - Treatment of disfluencies
- **Conclusion**

Conclusions

- Dialogue is an instance of *joint activity* between participants
- Three selected topics:
  - Can we parse dialogue *incrementally*?
  - Can we *learn* how to generate feedback?
  - How should we treat *disfluencies*?
- If you would like to collaborate with me on some of these aspects, let me know ;-)