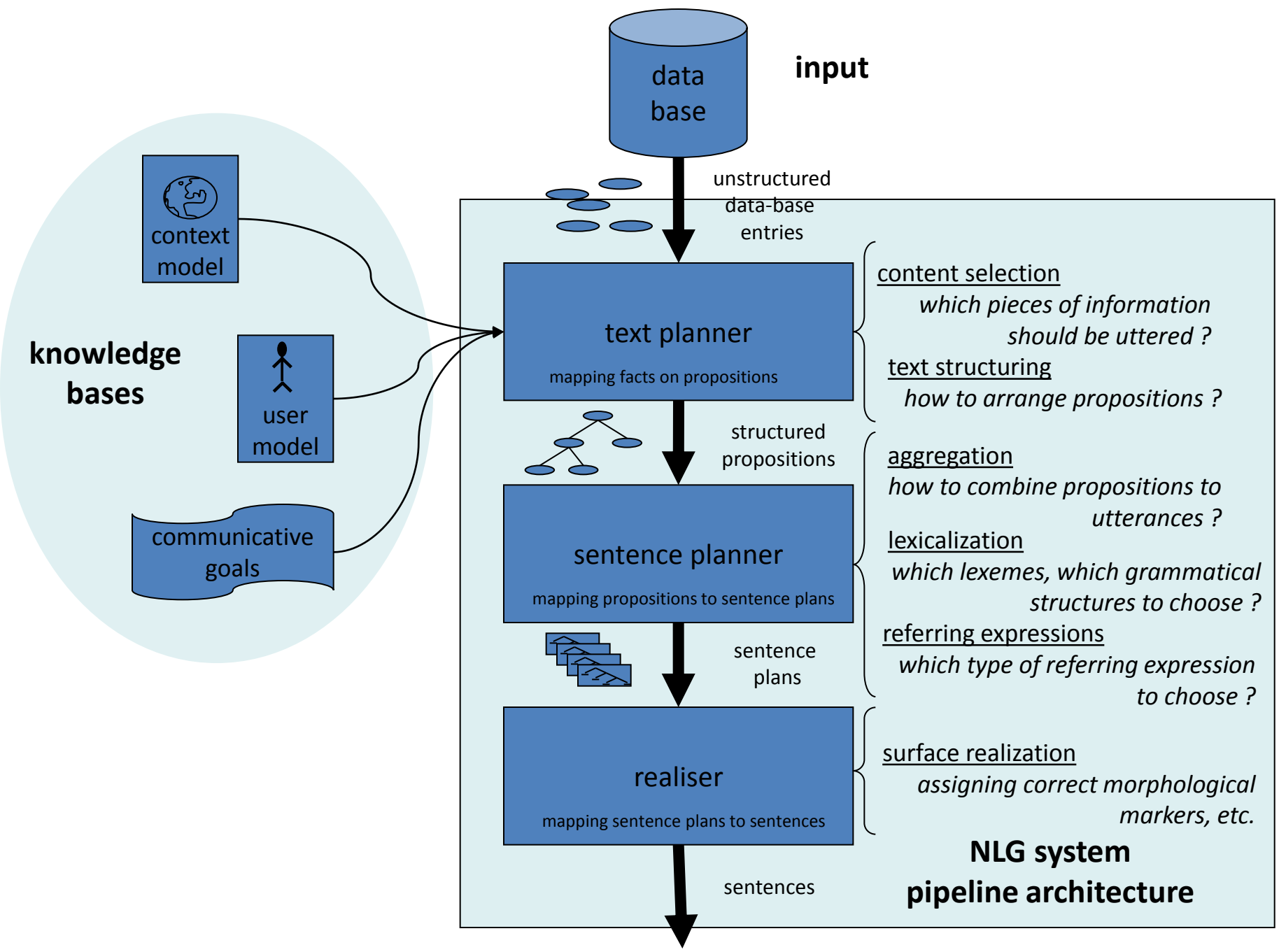
STS for NLG

Christian Chiarcos
chiarcos@uni-potsdam.de

Natural Language Generation

- Natural Language Generation (NLG) (...) is a subfield of artificial intelligence and computational linguistics that is concerned with building computer software systems that can produce meaningful texts in English or other human languages from some underlying non-linguistic representation of information.

Reiter & Dale 2000



Lorem ipsum dolor sit amet, consectetur adipisici elit ...

NLG applications

- generating text from large bodies of numerical data
  - weather reports (Belz 2008)
- generating text from a large knowledge bases
  - museum guide (O'Donnell et al. 2001)

- interactive hypertext
  - book recommendations (Chiarcos & Stede 2004)
    - taking the information status of the addressee into account
- user-tailored
  - BabyTalk (Gatt et al. 2009)
    - automatically generated medical reports for nurses/doctors (informative) and parents (affective)

- informative, instructional or persuasive texts

NLG evaluation: human

- task-oriented evaluation
  - measure impact on end user, e.g., mistakes (for an instructional text, Young 1999)
- human ratings and judgements
  - expert ratings according to criteria like coherence and (linguistic) quality (Lester and Porter 1997)
- expensive and time-consuming

NLG evaluation: automated

- evaluation by comparison with human written text
  - i.e., texts written by experts from the same data
    - or (in combination with a parser) corpus regeneration (Cahill and van Genabith 2006)
  - cheap, fast, repeatable (if we have the corpus)

NLG evaluation: automated

- n-gram metrics
  - BLEU (Papineni et al. 2002), from MT
  - ROUGE (Lin and Hovy 2003), from Summarization
- concerns
  - cannot capture higher-level information (e.g., information structure, Scott and Moore 2007)

  => evaluate correlation with human judgements (Reiter and Belz 2009)

NLG evaluation: automated vs. human

- Belz & Reiter (2009)
  - weather reports
  - human: experts and non-experts
  - automated: BLUE, ROGUE
  - criteria
    - „clarity and readability“ (= linguistic quality)
    - „accuracy and appropriateness“ (= content quality)

NLG evaluation: automated vs. human

- Belz & Reiter (2009)
  - significant correlations only with clarity, but not with accuracy
- strong influence on the design of subsequent NLG shared tasks
  - focus on task-based evaluation
    - GIVE, GIVE-2 (Giving Instructions in Virtual Environments)
    - GRUVE (Generating Route descriptions in Virtual Environments)
  - automated metrics mostly for the evaluation of surface realization
    - Surface realization challenge (BLUE, ROUGE, METEOR*)

* METEOR is a simple semantic metric using lexical similarity (synonyms)

NLG evaluation vs. STS

- Automated evaluation would benefit strongly from STS
  - automated, content-sensitive metrics are still an open research question in NLG
- NLG provides particularly strong motivation to include discourse in STS
  - unlike summarization and MT, we cannot just keep an existing structure