SDS Aplications
- Speech-to-speech translation -

Anca Burducea

May 28, 2015

S2S Translation

Three independent tasks:
$S_s \rightarrow T_s \rightarrow T_t \rightarrow S_t$

S_s = speech source
T_s = text source
T_t = text target
S_t = speech target

S2S Translation

$S_s \rightarrow T_s$ = ASR
$T_s \rightarrow T_t$ = MT
$T_t \rightarrow S_t$ = TTS

S_s = speech source
T_s = text source
T_t = text target
S_t = speech target

↓

Wo ist das nächste Hotel?

↓

Where is the nearest hotel?

↓

S2S Translation

$S_s \rightarrow T_s$ = ASR – WER
$T_s \rightarrow T_t$ = MT – BLEU
$T_t \rightarrow S_t$ = TTS – subjective listening tests

S_s = speech source
T_s = text source
T_t = text target
S_t = speech target

↓

Wo ist das nächste Hotel?

↓

Where is the nearest hotel?

↓

S2S Translation - Issues

- error propagation
- not using context in the downstream process

Annotations of Speech

A lot of context annotation on speech

- dialog act (DA) tags
- semantic annotation
- pitch prominence
- emphasis
- contrast
- emotion
- speaker segmentation

Sridhar 2013

Enrich S2S translations using contextual information!

Sridhar 2013

Enrich S2S translations using contextual information!

- DA tags
- prosodic word prominence

Sridhar 2013

Enrich S2S translations using contextual information!

- DA tags
- prosodic word prominence

Purpose:

- resolve ambiguities
  - wir *haben* noch $\rightarrow$ we *still* have
  - wir haben *noch* $\rightarrow$ we have *another*

Sridhar 2013

Enrich S2S translations using contextual information!

- DA tags
- prosodic word prominence

Purpose:

- resolve ambiguities
  - wir *haben* noch $\rightarrow$ we *still* have
  - wir haben *noch* $\rightarrow$ we have *another*
- enrich target speech with prosody (intonation, emotion) from source speech

Sridhar 2013

S_s = speech source
T_s = text source
T_t = text target
S_t = speech target
L_s **= enriched source** = text source + context labels
L_t **= enriched target** = text target + context labels

Sridhar 2013

S_s = speech source
T_s = text source
T_t = text target
S_t = speech target
L_s = **enriched source** = text source + context labels
L_t = **enriched target** = text target + context labels

$$S_t^* = \arg\max_{S_t} P(S_t|S_s)$$

$$\max_{S_t} P(S_t|S_s) \approx \max_{S_t} P(S_t|T_t^*, L_t^*, L_s^*). \max_{T_t, L_t} P(T_t, L_t|T_s^*, L_s^*). \max_{L_s} P(L_s|T_s^*, S_s). \max_{T_s} P(T_s|S_s) \quad (4)$$

Augmented Text-to-Speech | **Enriched Machine Translation** | **Rich Annotation** | **Speech Recognition**

Sridhar 2013

Data

- train MaxEnt classifier for
  - DA tagging: statement, acknowledgment, abandoned, agreement, question, appreciation, other – 82.9%
  - prosodic prominence: accent, no-accent – 78.5%

Sridhar 2013

Data

- train MaxEnt classifier for
  - DA tagging: statement, acknowledgment, abandoned, agreement, question, appreciation, other – 82.9%
  - prosodic prominence: accent, no-accent – 78.5%
- tested on three parallel corpora: Farsi-English, Japanese-English, Chinese-English

Sridhar 2013

Improve translation model using source language enrichment:

- bag-of-words model

$$T_t^* = \arg\max_{T_t} P(T_t|T_s, L_s)$$

$$= \arg\max_{T_t} \frac{P(T_s|T_t, L_s).P(T_t|L_s)}{P(T_s|L_s)}$$

$$= \arg\max_{T_t} P(T_s|T_t, L_s).P(T_t|L_s)$$

-
- reorder words according to target language model

Sridhar 2013

Improve translation model using source language enrichment:

- bag-of-words model

$$T_t^* = \arg\max_{T_t} P(T_t|T_s, L_s)$$

$$= \arg\max_{T_t} \frac{P(T_s|T_t, L_s).P(T_t|L_s)}{P(T_s|L_s)}$$

$$= \arg\max_{T_t} P(T_s|T_t, L_s).P(T_t|L_s)$$

-
- reorder words according to target language model

Improve translation model using target language enrichment

- factored model: word is translated into (word, pitch accent)

Sridhar 2013 - Results

DA tags

- question(YN, WH, open), acknowledgement $\rightarrow$ significant improvement
- statement $\rightarrow$ no significant improvement

Sridhar 2013 - Results

DA tags

- question(YN, WH, open), acknowledgement $\rightarrow$ significant improvement
- statement $\rightarrow$ no significant improvement

Prosody

- improved prosodic accuracy of target speech
- lexical selection accuracy no affected (same BLEU)

Sridhar 2013 - Results

DA tags

- question(YN, WH, open), acknowledgement $\rightarrow$ significant improvement
- statement $\rightarrow$ no significant improvement

Prosody

- improved prosodic accuracy of target speech
- lexical selection accuracy no affected (same BLEU)

Conclusion:
"the real benefits of such a scheme would be manifested through human evaluations. We are currently working on conducting subjective evaluations."

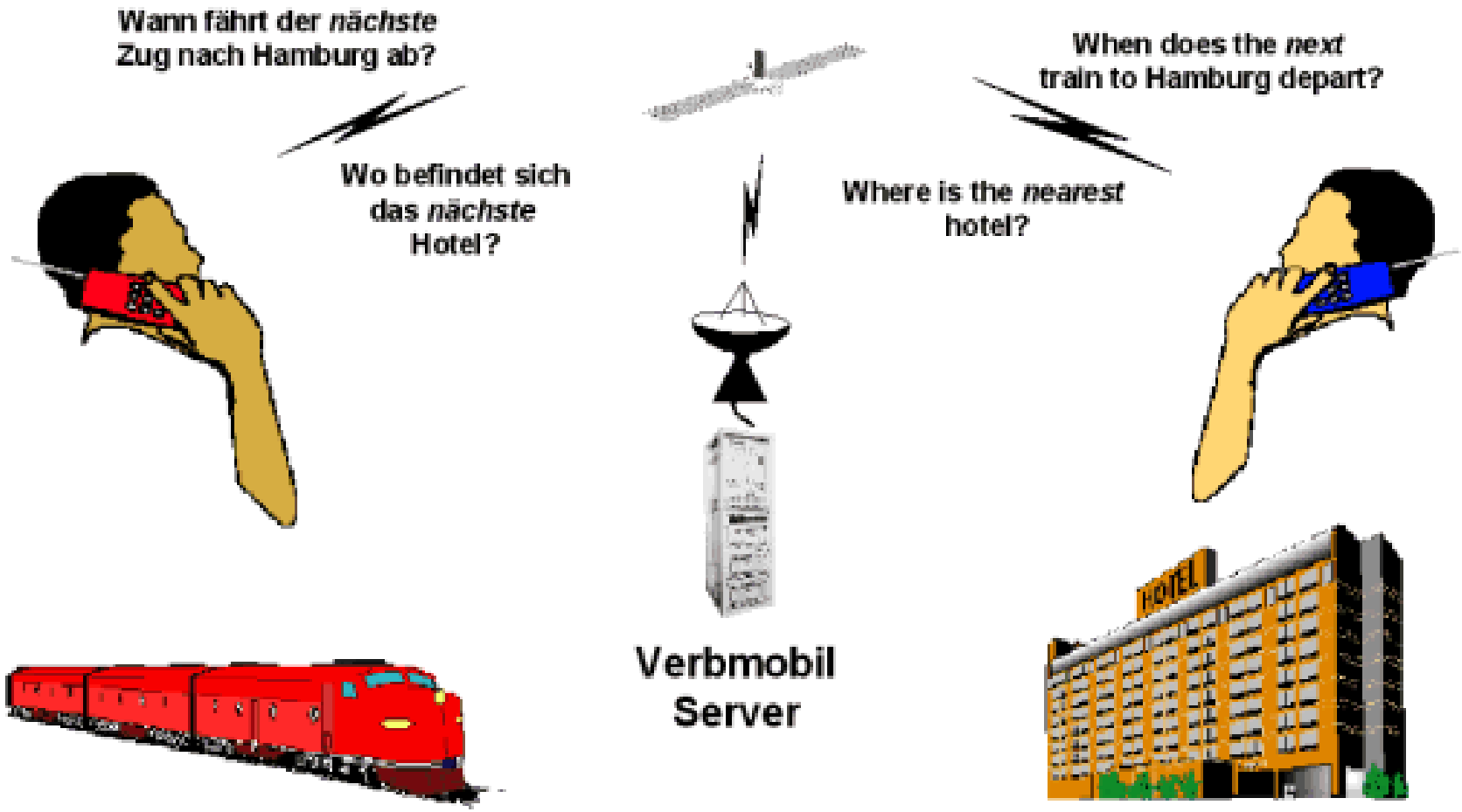
VERBMOBIL

- German S2S system developed between 1993-2000
- "verbal communication with foreign interlocutors in mobile situations"
- "Verbmobil is the first speech-only dialog translation system"
- bidirectional translations for German, English, Japanese
- business-oriented domains:
  1. appointment scheduling
  2. travel planning
  3. remote PC maintenance

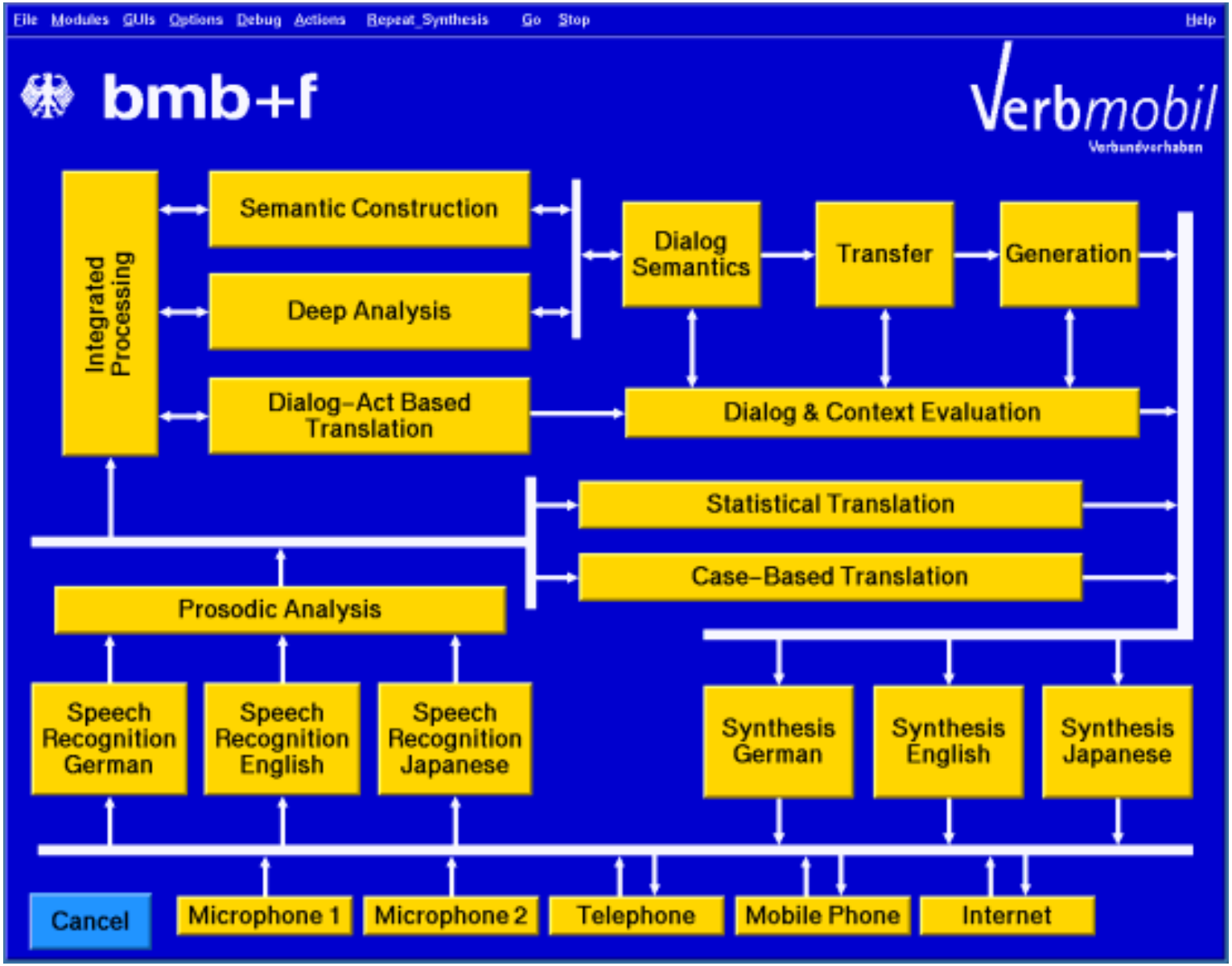
VERBMOBIL features

- context-sensitive translations
  e.g. GER nachste $\rightarrow$ ENG next (train) or nearest (hotel)
- prosody
  e.g. "wir *haben* noch" vs. "wir haben *noch*"
- domain knowlege: it knows "things about the topic being discussed"
- dialog memory: it knows "things that were communicated earlier"
- disfluencies management:
  1. filters out simple disfluencies ("ahh", "umm")
  2. remove reparandum

VERBMOBIL - Disambiguation



VERBMOBIL - Control Panel



Demo: