I. Introduction

Background: In highly interactive settings, Incremental Spoken Dialogue Systems are preferred over non-incremental systems as they read faster and more naturally (Liu et al., 2005).

- need to build up syntactic and semantic structure on the fly while the user is still speaking.
- needs theory-neutral assessment of the quality of incrementally built semantic structure.

Research Goals

- present generic measures to evaluate incremental semantics construction
- focus on measuring the incrementality (not necessarily the quality of the non-incremental result)
- show and analyze the performance of our specific module on a specific corpus

II. Incremental Semantics Construction

- Approaches vary by strictness of incrementality (Shah, 2004) vs. ambiguity of structures
- classical trade-off: slow and precise vs. quick and vague

Previous Work

Aist et al. (2006) and Rücker et al. (2006) generate (partial) hypotheses once semantics can be constrained to a small set.

“move a large triangle to…”

now(X, Y) now(triangle, Y)

Schuler (2004), Brick and Scheutz (2007), and others generate (2D) structures that might possibly match in the future.

“move a large triangle to…”


Neither of the previous works evaluates the incrementality of the semantic interpretation on a corpus. Partial evaluation is intrinsic to the used semantics construction mechanisms.

III. Evaluating Incremental Semantics Construction

There is no incremental gold-standard!

- in non-incremental evaluation, we can just compare to the human annotated gold-standard
- there is no (cannot be?) annotation of what should be known up to a certain word in the utterance

All in all, we want the final result (or parts thereof) as soon as possible.

Measures

The measures we propose relate relative position in the utterance and comparison to the non-incremental gold.

Also, we differentiate both partial and complete success.

- first correct element (FCE): When is the first element of the representation correct?
- first correctly-filled representation (FCR): When (in percent) do we first match the gold-standard?
- first finally correctly-filled representation (FFR): When does the correct representation not change anymore?
- degree of correctness (DC) at a certain time: The percentage of elements correctly filed on average.

IV. Application

Domain


- manually transcribed with a 5-slot frame-semantics
- 500 utterances: 100 for grammar development, 400 in the evaluation

332 test utterances (68 utterances had completely empty frames)
- we distinguish 171 short utterances (<10 words) and 161 long utterances (>10 words)

- interpretation can often be completed before the utterance is over
- this is especially true for long utterances.
- parts of the interpretation are already first correct (FCE) in the beginning of an utterance
- this could be e.g. used to prepare possible system responses.
- high degree of correctness increase with time, low decrease
- considerable knowledge after only 40% of the utterance
- relative stability in DC between 40% to 80% of the utterance
- the first few and the final word in the utterance are most important.

V. Discussion

We have defined measures to evaluate the incrementality of semantic components against a non-incremental gold-standard:

- measures seem to capture meaningful aspects of incremental semantic intepretation.
- measures are generic enough and allow to compare components with differing approaches to semantic interpretation.

We have evaluated our semantic component and found that incremental semantic interpretation is worthwhile:

- considerable knowledge only in parts of the utterance available.

Our results are obviously limited to corpus and semantics:

- standardized corpora and annotations are needed to compare different approaches.

Bibliography


Further Information

Please contact us via e-mail at
atterer@uni-potsdam.de

This work is funded by a DFG grant in the Emmy Noether programme.