CENTRE FOR LANGUAGE TECHNOLOGY
MACQUARIE UNIVERSITY - SYDNEY

Automatic Evaluation of Referring Expression Generation is Possible

Jette Viethen
jviethen@ics.mq.edu.au

Preliminaries

Why do we want shared evaluation?

- It has benefited other fields.
- To learn about evaluation techniques.
- For fun.
- To provide resources.
- To measure and ensure progress.

What do we want to evaluate?

- Applications or NLG subtasks?

How do we want to evaluate?

- Competitive – Comparative – Collaborative
- Automatic or human evaluation?

Agenda

- Yes, REG is still mainly focussed on distinguishing initial reference.
- Taking an optimistic look at the 5 main challenges for REG evaluation:
  - Defining Gold Standards
  - Output Expectations
  - Parameters
  - A wide field with few players
  - Input Representation

Defining Gold Standards

- For automatic evaluation, we need gold standard corpora to which to compare system output.
- There is never just one correct answer in NLG.
  - Every object can be described in many acceptable ways.
- A gold standard for REG needs to contain “all” acceptable descriptions for each object to be fair.
- The TUNA corpus looks like the right point.

Output Expectations

Quantity: Are we content with only one solution?

- Evaluate one description per object from each system – for now.
- Maybe later allow multiple entries.

Quality: What is a "good" referring expression?

- Get people to rank different descriptions for the same object.
- Assess usability by success rate and time.
- Many factors make it hard to assess one subtask.

Linguistic Level: From content determination to surface realisation.

- Concentrate on content determination – for now.

Parameters

- Most REG systems take one or a number of parameters.
- Very fine grained parameters allow the engineering of virtually any desired output.

What do we want to evaluate?

- The theoretical capacity of a system: Parameter part of the system and not be switched during an evaluation.
- The actual execution of the task: Automatic determination of the best parameter settings for describing a certain object.

A wide field with few players

- We need to use our human resources wisely!
- REG has many people working on it and is well defined.
- However, people are working on many sub-problems and domains.
- Concentrating on one competitive task would divert attention from other important areas.
- The evaluation corpus needs to cover a number of domains and be subdividable into types of referring expressions.

Input Representation

Counting from infinity to infinity...

- Highly dependent on application domain.
- Tightly intertwined with algorithm design.
- Let everyone choose their own representation.
  - Representation is part of the system.
  - Challenge of finding the same properties that people used.
- Agree on a common underlying knowledge base.
  - Based on properties and relations used in the corpus.
  - Input representation and algorithm design can be detangled.

Summary

To get started with automatic evaluation for REG:

- Build a corpus containing as many “good” REs per object as possible.
- Get human rankings for the REs in the corpus.
- Concentrate on a low linguistic level for now.
- Treat parameter settings as part of the algorithm.
- Include many different kinds of REs in the corpus.
- Don‘t compete on one task. Share resources.
- Standardise the underlying knowledge base.