The Attribute Selection for Generation of Referring Expressions Challenge

Organised by Anja Belz, Albert Gatt, Ehud Reiter and Jette Viethen

Introduction

A Shared Task Evaluation Challenge (STEC) is an exercise where typically participants are asked to develop systems that perform a Natural Language Processing (NLP) task, based on a data set (training data) which specifies example inputs and outputs for this task. Systems submitted to the STEC are then tested on test data which is similar to the training data, using a metric specified in the challenge.

STECs are common in many areas of NLP, including Machine Translation, but STECs have not previously been tried in the area of Natural Language Generation (NLG). However, interest in comparative evaluation has steadily been growing in NLG: in July 2006, Anja Belz and Robert Dale organised the Special Session on Sharing Data and Comparative Evaluation at INLG 2006 (`http://www.nltg.brighton.ac.uk/home/Anja.Belz/inlg06-specsess.html`), and as a result of discussions there, in April 2007 Michael White and Robert Dale organised a dedicated two-day Workshop on Shared Tasks and Comparative Evaluation in NLG (`http://www.ling.ohio-state.edu/m̃white/nlgeval07/`), with the support of the National Science Foundation. Working groups at this workshop discussed three areas for NLG STECs: text-to-text applications, instruction giving in virtual worlds, and generating referring expressions (GRE); a fourth group explored general desiderata and concerns for NLG STECs.

The GRE working group decided to go ahead and organise a pilot STEC in the GRE area, which focused on the attribute selection task (the most commonly researched GRE task) and was based on the Aberdeen TUNA corpus (`http://www.csd.abdn.ac.uk/research/tuna/corpus/`); this became the Attribute Selection for Generation of Referring Expressions Challenge. The pilot STEC had two main objectives: gauge how interested other NLG researchers were in participating in a STEC, and try out our ideas about how to best organise an NLG STEC.

Organising this event required us to make numerous decisions, including

- What sort of submissions to invite: in addition to the shared task proper, we decided to invite submissions proposing new tasks and/or new evaluation techniques.
- What data to release: we decided to pre-process the raw TUNA corpus data in order to simplify and rationalise it.
- How to evaluate submissions: we decided to use the Dice metric and also to perform a task-based evaluation with human readers.
- How to organise the STEC: we decided to use a similar organisation to STECs in other areas of NLP.

Looking back, some of these decisions may not have been ideal; for example perhaps we over-simplified the TUNA corpus. But certainly we have learned a great deal about organising an NLG STEC, which was one of our main objectives in holding this pilot event.

We were very pleased with the level of interest shown; we received 19 individual registrations for the shared-task track, and 13 of these formed six teams which submitted 22 systems by the deadline. We also received one submission in the evaluation methods track. This shows that there is considerable interest in STECs in the NLG community.

Encouraged by the level of interest shown in the community and the lessons we have learned, we are planning to organise another STEC in the GRE area in 2008, which will be larger and offer multiple shared tasks.

Our thanks to Robert Dale and Michael White for organising the initial STEC workshop; to Irene Langkilde-Geary and the other faculty and staff at Brighton University for their help with the evaluations, and to all of the people who participated in this event.

September 2007 Ehud Reiter, Anja Belz, Albert Gatt and Jette Viethen